Diversité, évolution et écologie virale : des communautés aux génotypes. Analyse bioinformatique de métagénomes viraux

Simon Roux

UNIVERSITE BLAISE PASCAL — UNIVERSITE D'AUVERGNE

N° D.U. : 2380

ECOLE DOCTORALE SCIENCES DE LA VIE , SANTÉ, AGRONOMIE, ENVIRONNEMENT

N° d'ordre : 615

Thèse présentée à l'Université Blaise Pascal

pour l'obtention du grade de Docteur d'Université (Spécialité : Écologie)

Diversité, évolution, et écologie virale: des communautés aux génotypes

Analyse bioinformatique de métagénomes viraux

Simon Roux

Thèse dirigée par **Didier Debroas** et **François Enault**

Soutenue publiquement le **03 octobre 2013** devant le jury composé de :

Télesphore SIME-NGANDO, Directeur de Recherche, CNRS, Aubière (Président)

Denis LE PASLIER, Directeur de Recherche, CNRS, Genoscope, Evry (Rapporteur)

Hervé MOREAU, Directeur de Recherche, CNRS, Banyuls (Rapporteur)

Pascal HINGAMP, Maître de Conférence, Université de la Méditérannée (Examinateur)

Laboratoire Micro-organismes : Génome et Environnement

Remerciements

Après plus de trois années, un grand nombre de personnes ont participé directement ou indirectement à ce travail, que ce soit par des collaborations, des discussions scientifiques, ou au contraire des moments très loin de la science indispensables pour ne pas sombrer dans l'obsession des virus. Je prie donc tous ceux que j'aurais l'impolitesse et le malheur d'oublier de bien vouloir m'excuser. et à tous et à toutes, un grand merci !

Bien évidemment, il me faut remercier en premier lieu Didier et François, sans qui ce travail n'aurait pu exister. Merci pour tout, en particulier les conseils, la disponibilité, et la liberté que j'ai pu avoir au fur et à mesure de l'avancement de la thèse.

Je souhaiterais également exprimer ma gratitude à Messieurs Hervé Moreau et Denis Le Paslier, qui m'ont fait l'honneur d'accepter d'être rapporteurs de cette thèse, ainsi que Mr Pascal Hingamp pour avoir accepté un rôle d'examinateur au sein de ce jury.

Merci également à Christian Amblard et Télesphore Sime-Ngando pour l'accueil au sein du LMGE (Laboratoire Micro-organismes Génome et Environnement, UMR CNRS 6023), et pour m'avoir donné les moyens matériels de réaliser ces travaux, notamment en soutenant le développement du serveur Metavir. Merci à la direction générale de l'armement et ses correspondants, Gilles Vergnaud, Emmanuelle Guillot-Combe et Catherine Eng, ainsi qu'à l'agence nationale de la recherche (projet SENDEFO) pour avoir soutenu financièrement cette thèse.

Merci à l'ensemble de l'équipe MEB (Microbiologie de l'Environnement et Bioinformatique) pour ces trois années ou, grâce à vous, venir au travail était tout simplement agréable. Merci en particulier à Viviane Ravet, Agnès Vellet et Emilie Duffaud, sans qui je n'aurais tout simplement pas eu de données à analyser. Je tiens également à exprimer toute ma gratitude à Gisèle Bronner, pour m'avoir inculquer un peu plus de rigueur, et pour toutes les discussions qui ont considérablement fait avancer ma vision de la biologie et de l'évolution. Merci à tous les responsables de modules qui m'ont accueilli durant mon service de monitorat (notamment Isabelle Jouan), qui m'ont permis de réaliser mes premiers pas dans l'enseignement dans des conditions exceptionnelles. Enfin, merci à Jean-Christophe Charvy pour les différentes aides techniques, mais aussi pour le soutien au quotidien.

Je tiens également à remercier l'équipe VMM (Virus et Métabolisme Microbien) et en particulier Agnès Robin, Jonathan Colombet et Stéphanie Palesse pour le soutien et toutes les informations fournies. Merci à Antoine Mahul et au CRRI (Centre Régional de Ressources Informatiques) pour le support technique et l'aide précieuse apportée, en particulier pour la mise en place du serveur Metavir.

Toute ma gratitude à Patrick Forterre, Jean-Luc Bailly, Christelle Desnues et Mickaël Desvaux pour avoir accepté de participer à mes différents comités de thèse, et pour leur précieux conseils qu'ils ont pu m'apporter à cette occasion.

Merci à Jean-François Humbert, Julie Leloup, Stéphane Pesce, Noémie Pascault, David Prangishvili, Yvan Bettarel et Jean-Christophe Auguet pour leur accueil chaleureux, et tous les conseils et discussions à l'occasion des différents projets communs. Dans ce cadre, je me dois de remercier tout spécialement Mart Krupovic, pour les différentes collaborations qui ont pour moi été extrêmement enrichissantes, et pour sa grande disponibilité. Merci également à l'ensemble des membres du laboratoire Information Génomique et Structurale, et en particulier à Jean-Michel Claverie et Hiroyuki Ogata pour leurs nombreux conseils et avis.

Merci à mes deux stagiaires Axel Poulet et Jeremy Tournayre. En particulier, merci Axel pour tout le travail d'analyse de séquence effectué, et merci Jeremy pour la masse de travail considérable abattue dans les différents stages. Il est clair que les différentes études n'auraient pu être réalisée sans votre aide. Merci également à Daniel Vaulot, Sébastien Personnic, Aurélien Bernard et Michaël Faubladier pour leur participation dans les différentes études réalisées.

J'ai une grande pensée pour l'ensemble de la promo 2010 du master AMDSV : Seb, Jo, Nils Mathieu, grâce à qui les deux années de Master ont été mémorables. Merci notamment à Seb pour le sujet sur les viromes, je n'oublie pas que c'est toi qui l'avait commencé ! Et puis on avait bien dit que ça m'emmènerait sur la côte ouest, on était juste à 2500 km trop au nord.. ! Merci également à tous les historiens pour les pique-nique du midi au jardin Lecoq, coupure bien agréable dans la journée. Thomas et Laurie, merci pour le soutien, la présence constante à nos côtés, et tous les moments ou enfin, il était possible de penser à autre chose qu'à la biologie ! Un grand merci également à Pierre et Laura, notamment pour tous ces jeudis soirs qui, s'ils rendaient les vendredi matin parfois un peu compliqués, rendaient toute la semaine tellement plus facile. Vous êtes tous adorables, et avez été d'un grand soutien !

Enfin, merci à toute ma famille, mes parents pour tous les conseils, l'incitation à être curieux de tout, et la confiance témoignée, et tous mes frères et sœurs, neveux et nièces, cousins et cousines. Et pour terminer ces remerciements, je manque de mots pour exprimer l'étendue de ma gratitude envers Pauline. Ton soutien, ta patience et ta confiance durant toutes ces années, depuis la licence de biologie vers laquelle que tu m'as poussé, jusqu'à ces trois années de doctorat, ont été exceptionnels. J'ai tout simplement énormément de chance que tu ais été et que tu sois encore à mes côtés.

Résumé

Les virus sont omniprésents dans la biosphère et infectent vraisemblablement l'ensemble des êtres vivants. Au sein des écosystèmes, ils ont ainsi un impact sur la diversité des populations microbiennes, l'évolution des génomes de ces populations, et directement ou indirectement sur les cycles biogéochimiques majeurs. Leur caractère protéiforme et l'absence de marqueur unique (tant génétique que physique) font toutefois de l'exploration de la diversité virale une tâche complexe, de telle sorte que nos connaissances sur ces communautés virales environnementales sont encore très limitées.

La métagénomique, ou séquençage massif et aléatoire de fragments nucléotidiques extraits d'un prélèvement, offre un point de vue unique sur les génomes viraux. Ce type d'approche, récemment développé, a ainsi mis en évidence la richesse extraordinaire des populations virales environnementales, tant du point de vue des gènes que des génotypes.

C'est dans ce cadre de l'étude des communautés virales de l'environnement par métagénomique que se sont inscrits les travaux de cette thèse, organisée autour de quatre axes principaux :

- Le développement de nouvelles méthodes d'analyses adaptées aux spécificités des génomes et métagénomes viraux par la mise en place du serveur web Metavir, premier serveur dédié à l'analyse des viromes. Proposant aujourd'hui un ensemble cohérent d'outils pour différents types de viromes, Metavir compte plus de 300 utilisateurs pour plus de 2000 viromes analysés.
- Le potentiel fonctionnel des génomes viraux a pu être approché par l'étude conjointe d'un ensemble de viromes. Après une analyse rigoureuse des contaminations potentielles, nous avons pu confirmer que les génomes viraux comprenaient un ensemble limité mais non négligeable de gènes associés au métabolisme cellulaire. La plupart des virus agissent ainsi certainement directement sur le métabolisme de la cellule hôte durant l'infection.
- La prépondérance des paramètres environnementaux, et particulièrement de la salinité, en tant que facteurs structurant les communautés virales aquatiques a également pu être mise en avant. La distance géographique entre prélèvements semble n'avoir qu'une influence secondaire, confirmant la capacité importante de dispersion des capsides virales. Une adaptation locale semble toutefois exister dans certains cas, notamment en cas de compétition importante entre les résistances développées par les hôtes et les capacités d'infection des virus.
- Enfin, différentes familles de petits virus à ADN simple brin ont pu être caractérisées par une méta-analyse de viromes. Leur apparente simplicité a ainsi révélé des mécanismes d'évolution plus complexes que prévus, impliquant différents cycles et capacités de transfert de gènes jusqu'ici plutôt considérés comme l'apanage des virus à ADN double brin, et remettant en cause les séparations admises entre les différents groupes de virus sur la base de la nature de leur génome.

En permettant une étude depuis l'échelle de la communauté jusqu'à des génotypes spécifiques, les viromes constituent des outils de choix pour caractériser la diversité virale, appréhender les différents facteurs régulant ces communautés, et ainsi mieux comprendre la place des virus dans la biosphère. De plus, ces études ont confirmé l'existence d'interactions étroites entre virus et organismes cellulaires, ces interactions semblant nombreuses, multiples dans leurs natures et conséquences, et présentes tout au long de l'histoire du vivant. Ces nouvelles connaissances apportées par l'analyse de viromes permettent donc d'aborder certaines questions fondamentales concernant l'origine des grandes innovations évolutives ou le fonctionnement global des écosystèmes.

Mots-clés : Virus / Métagénomique / Bioinformatique / Écologie / Génomique / Évolution.

Abstract

Viruses likely infect every organism on Earth (in some cases even other viruses!), and represent vast morphological and genetic diversity. Not surprisingly given their numerical dominance, viruses significantly impact ecosystems through regulating microbial populations, driving major biogeochemical cycles, and shaping the evolution of hosts genomes. However, our understanding of viruses in nature is primitive, especially because the majority of environmental viral genomes remains uncharacterized.

Metagenomics (*i.e.* random and massive sequencing of genomic fragments isolated from a sample) applied to encapsidated genetic templates provides a unique perspective on the viral pangenome. The first viral metagenomes (or viromes) generated entire sets of new questions about viral diversity, especially concerning their genetic and species richness.

This work was set within this frame of viral diversity study through metagenomics, and organized into four main themes :

- The development of bioinformatics tools adapted to the specific features of viral genomes and metagenomes led to the release of Metavir, the first web server dedicated to virome analysis. Providing a comprehensive set of connected tools, Metavir has now been used by more than 300 users in the analysis of more than 2000 viromes.
- The functions encoded within viral genomes were for the first time thoroughly examined, following a rigorous examination of a set of published viromes toward contamination by cellular DNA. A new picture of the viral functional potential could thus be drawn, which confirmed that the range of cellular functions encoded in viral genomes is wider than the one retrieved from the complete genomes currently available, though not as great as previously estimated.
- The study of the aquatic viral metagenomes also revealed the importance of salinity in the distribution of viral communities across the globe. The ubiquitous distribution of most viral genotypes confirmed that viral particles seem to be able to move across any distance on Earth. Viruses are thus likely selected based on factors such as the presence of their host in the samples and the competition with other parasites, which can still drive local adaptations.
- Finally, viromes were used to better characterize the diversity of different ssDNA viral families. Despite their small size and relative simplicity, these viruses were found to harbor unexpectedly complex cycles and evolutionary mechanisms, in particular a great potential of recombination and gene transfer. Overall, the new genomes assembled from viromes notably challenge the separation between viruses based on the nature of their genome.

Eventually, as illustrated by these different works and analyses, viromes are unique and extremely powerful tool to assess and characterize viral genetic diversity. Moreover, considering the tight links between viral and cellular worlds, insights into the viral communities provided by metagenomics make it possible to address fundamental questions such as the origin of important evolutive innovations or the functioning of ecosystems, so that these results are of interest for the whole field of biology.

Keywords: Virus / Metagenomics / Bioinformatics / Ecology / Genomic / Evolution

Table des matières

Liste des figures et tableaux

Cette liste ne tient pas compte des figures et tableaux contenus dans les articles.

Introduction et état de l'art

Chapitre I- Développement d'outils bioinformatiques dédiés à l'analyse de viromes : le serveur web Metavir

Chapitre II – Potentiel fonctionnel des génomes viraux

Chapitre III – Facteurs structurants des communautés virales aquatiques

Chapitre IV – Virus à ADN simple brin : diversité et mécanismes d'évolution

Conclusion

Liste des abréviations

ADNr : ADN ribosomal (séquence génomique codant pour un ARN ribosomal)

COG : Cluster of Orthologous Groups (Clusters de groupes de gènes orthologues)

CRISPR : Clustered Regularly Interspaced Short Palindromic Repeats (Cluster de courtes répétitions palindromiques régulièrement espacées)

DGGE :Denaturing Gradient Gel Electrophoresis (Electrophorèse sur gel en conditions dénaturantes)

FISH : Fluorescent *In Situ* Hybridization (Hydbridation in situ fluorescente)

GTA : Gene Transfer Agent (Agent de transfert de gènes)

kb : Kilobase

HTS: High-Throughput Sequencing (Séquençage à haut-débit)

ICTV : International Commitee on Taxonomy of Viruses (Comité international sur la taxonomie des virus)

LUCA : Last Universal Common Ancestor (Ancêtre commun universel le plus récent)

MDA : Multiple Displacement Amplification (Amplification par déplacements multiples de brin)

MDS : Multidimensional Scaling (Positionnement multidimensionnel)

MET : Microscopie Électronique à Transmission

Moron : More DNA (ADN supplémentaire)

NCBI : National Center for Biotechnology Information

NMDS : Non-Parametric Multidimensional Scaling (Positionnement multidimensionnel non paramétrique)

ORF : Open Reading Frame (Cadre de lecture ouvert)

ORFan : Orfan ORF (Cadre de lecture ouvert orphelin)

pb : Paire de bases

PCR : Polymerase Chain Reaction (Réaction en chaine par polymerase)

PFGE :Pulse-Field Gel Electrophoresis (Electrophorèse sur gel en champs pulsé)

POG : Phage Orthologous Groups (Groupes de gènes d'orthologues au sein des phages)

RAPD-PCR : Random Amplification of Polymorphism DNA (Amplifiction aléatoire d'ADN polymorphe)

RC-Rep : Rolling-circle réplication (Réplication par cercle roulant)

SCREVs : Small Circular Rep-Encoding Viruses (Petits virus circulaire encodant un gène RC-Rep)

Introduction et état de l'art

1. Virus en biologie – découverte, définition et premières classifications

Historiquement, le mot “virus” a tout d'abord désigné n'importe quel type d'agent infectieux, sans indication sur sa nature exacte. En latin, ce mot regroupe en effet les notions de “poison, venin, sécrétion, humeur et infection”, et désignera ainsi jusqu'au XIX^{e} siècle tout vecteur de contagion non caractérisé, avant d'être par la suite associé à un type particulier d'agent infectieux (Chastel, 1992).

Premières observations et caractérisations d' “agents filtrables infectieux”

Spéculations autour de la nature des virus

La première description d'un virus au sens contemporain du terme vient des travaux de Dimitri Ivanovski (1864 - 1920) portant sur une maladie survenant sur les plants de tabac (la mosaïque du tabac). Ce botaniste russe isolait des agents infectieux par l'utilisation de filtres de porcelaines (aussi appelés bougies de Chamberland) comme il était d'usage à la fin du XIX^{e} siècle. En 1892, il présenta à l'académie impériale des Sciences de Saint Petersbourg des résultats montrant que l'agent infectieux de la mosaïque du tabac était contenu dans la sève des plantes, et qu'il n'était pas retenu par les filtres Chamberland (Zaitlin, 1998).

Dans un premier temps, cet agent fut considéré comme étant soit une très petite bactérie, soit une toxine. Plusieurs travaux complémentaires réfutèrent toutefois ces hypothèses. Il fut ainsi démontré que la sève d'une plante infectée chauffée par une bougie gardait un pouvoir infectieux, ce qui n'était le plus souvent pas le cas lorsqu'il s'agissait d'agents bactériens. Les travaux du chimiste hollandais Martinus Beijerinck (1851 - 1931) démontrèrent que le pouvoir infectieux de la sève de plantes malades n'était pas diminué à la suite de dilutions multiples, ce qui aurait du se produire si l'agent avait été une toxine. Reproduisant les expériences d'Ivanoski en utilisant des filtres encore plus fins, Beijerinck conclut également qu'il ne pouvait s'agir d'une bactérie, et nomma pour la première fois ce pathogène “Virus” (Zaitlin, 1998). Dans le cadre de travaux sur le virus pied-main-bouche, Fredriech Loeffler (1852 - 1915) et Robert Koch (1843 - 1910) seront toutefois les premiers auteurs à clairement identifier le virus comme une petite particule non retenue par les filtres de Chamberland mais retenue par les filtres plus fins de Kitasato, là ou Ivanoski pensait avoir affaire à une sorte de petit microbe, et où Beijerinck considérait le virus comme une entité infectieuse liquide et non particulaire (Mahy, 2005).

Mise en lumière du monde viral et de son étendue

Au début du XX^{e} siècle, les découvertes d'éléments filtrables infectieux se multiplièrent, y compris chez les animaux (on peut citer les agents responsables de la peste équine ou de la myxomatose du lapin). Puis, le même type d'agent est identifié pour des maladies humaines comme la fièvre jaune (1902) et la rage (1903), ou animales comme la vaccine (ou “variole de la vache”, en 1905). Toutefois, la conceptualisation des virus en tant qu'entités de nature fondamentalement différente des bactéries est alors peu partagée par la communauté scientifique, et notamment retardée par la découverte des mycoplasmes (agents de la péripneumonie des bovidés), de minuscules microbes capables de passer les filtres les plus fins. L'idée généralement adoptée et formulée dans l'article décrivant la découverte des mycoplasmes est celle de l'existence de “microbes plus petits encore (que celui de la péripneumonie des bovidés), lesquels au lieu de rester en deçà des limites de la visibilité, comme c'est le cas pour celui-ci, seraient au-delà de ces limites ; en d'autres termes, on peut admettre qu'il existe des microbes invisibles pour les yeux de l'homme” (Roux, 1903).

Parallèlement à la description de ces virus eucaryotes, l'existence d' “agents filtrables infectieux” infectant cette fois les bactéries fut rapportée par Ernest Hanbury Hankin (1865 - 1939). Dès 1896, ce microbiologiste anglais notait qu'il semblait exister dans l'eau du Gange des agents passant les filtres de porcelaine et dotés d'un pouvoir anti-bactérien, notamment contre des pathogènes tels que l'agent du choléra (*Vibrio cholera*). La présence spécifique de virus infectant les bactéries (ou bactériophages) sera mise en lumière par les travaux de Frederick William Twort (1877 - 1950). En 1914, différents travaux utilisant des cultures bactériennes mèneront ce scientifique anglais à l'observation de zones translucides au sein des colonies en milieu liquide, qu'il identifia comme résultant de la destruction des cellules bactériennes. Il décrivit ainsi les principales caractéristiques des bactériophages en réussissant à infecter des colonies saines à partir de ces “taches”, et en démontrant que les agents infectieux passaient un filtre de porcelaine et nécessitaient la présence de la bactérie pour se développer. Il émit toutefois trois hypothèses sur la nature exacte de ces agents infectieux : soit il s'agissait de bactéries à un stade particulier de leur développement, soit d'une enzyme produite par les bactéries, soit d'un virus infectant ces bactéries (Duckworth, 1976).

Ces observations furent complétées par les travaux de Félix d'Hérelle en 1917, qui avait obtenu de manière indépendant des résultats très similaires. En effet, au cours de la culture de bactéries, ce franco-canadien observa et décrit les premières plages de lyse, et a l'intuition qu'il s'agit d'un agent infectant les bactéries. Il leur donnera d'ailleurs le premier le nom de “bactériophage”. Cependant, le débat sur la véritable nature de ces bactériophages durera plus de 15 ans après ces premières observations (Duckworth, 1976).

Identification de la nature spécifique des virus

Caractérisation structurale des virus

A la suite de ces premières observations de virus, les progrès techniques particulièrement en matière de culture cellulaire ont permis de mieux définir ces "agents filtrables infectieux". Ainsi, Arthur Edwin Boycott (1877 - 1938) dressait en 1928 un bilan des connaissances sur les virus conduisant à une définition étonnamment moderne de ces derniers (Chastel, 1997). Tout d'abord, il nota que les virus devaient être organisés sous forme particulaire, d'une taille réduite mais supérieure à celles des "sels simples", et indiquait même une estimation de l'ordre de 25 nm. Considérant leur petite taille et l'absence de détection de toute activité métabolique, A. Boycott indiqua qu'il était fort probable que les virus ne soient pas très élaborés au niveau des activités vitales. L'existence de réactions antigéniques aux infections virales témoignait de la nature probablement protéique des capsides virales, et plusieurs observations montraient que ces particules possédaient une grande stabilité, résistant au temps ainsi qu'à des agents dénaturants très variables. Enfin, une cellule vivante, de préférence jeune, était nécessaire pour leur reproduction.

Ainsi, les virus, jusque-là définis par un ensemble de traits négatifs (non retenus par les pores du filtres, incapables de vie autonome, impossible à observer), sont ici présentés plus positivement par la description de caractéristiques intrinsèques (taille, nature protéique, stabilité). Cette transition vers une description du monde viral à travers des caractères positifs sera poursuivie durant le deuxième tiers du XX^{e} siècle, jusqu'à l'adoption d'une nouvelle classification basée sur les caractéristiques structurales des virions.

C'est en effet en 1952 que la première structure d'un virus fut établie, avec la description du virus de la grippe, virus enveloppé, généralement rond, d'une taille variant entre 80 et 120 nm de diamètre (Hoyle, 1952). La structure tubulaire du virus de la mosaïque du tabac fut ensuite décrite en 1955 (Franklin, 1955). Enfin, la même année, l'assemblage spontané des capsides de virus de la mosaïque du tabac (complets et infectieux) à partir de l'ARN et des protéines dissociées fut démontré par Heinz Fraenkel-Conrat et Robley Williams (Fraenkel-Conrat & Williams, 1955). Ces différentes découvertes ne pouvaient que conduire à la séparation définitive entre les virus et les micro-organismes, formalisée notamment par Lwoff en 1957 : "***viruses are viruses****" (Lwoff, 1957)*.

Premières définitions et classifications

Au-delà de cette séparation, les virus étaient alors décrits comme infectieux, potentiellement pathogènes, de nature nucléo-protéique et possédant un seul type d'acide nucléique, incapables de croître et de se diviser de manière autonome, et sans système métabolique. Les premières classifications des virus furent ainsi élaborées dans les années 1960 (1962 puis 1966), et étaient alors basées sur la nature de l'acide nucléique génomique et les paramètres de la capside (type de symétrie, nombre de " capsomères ", diamètre, virus nus ou enveloppés). Ces classifications seront reprises et complétées par la **classification de Baltimore** (1971 ; Figure In.1) prenant en compte le mode de réplication en plus de la nature du génome, les virus étant ainsi regroupés au sein de sept classes différentes (Baltimore, 1971).

Le dernier tiers du XX^{e} siècle sera un véritable âge d'or pour l'exploration du monde viral, avec la description et l'observation d'une extraordinaire diversité de formes, de modes de réplication et d'interactions avec l'hôte. Certaines découvertes vont éprouver les limites mêmes de la définition du virus, à l'image du virus Epstein-Barr et de son association potentielle aux maladies dites "lentes" (comme le lymphome d'Hodgkin), les virus oncogènes et notamment le virus du sarcome de Rous, les structures de type viroïde (particules composées d'une seule molécule d'ARN et sans capside, décrites pour la première fois en 1971 par Theodor O Diener), la description de virus géants aussi complexes que les plus

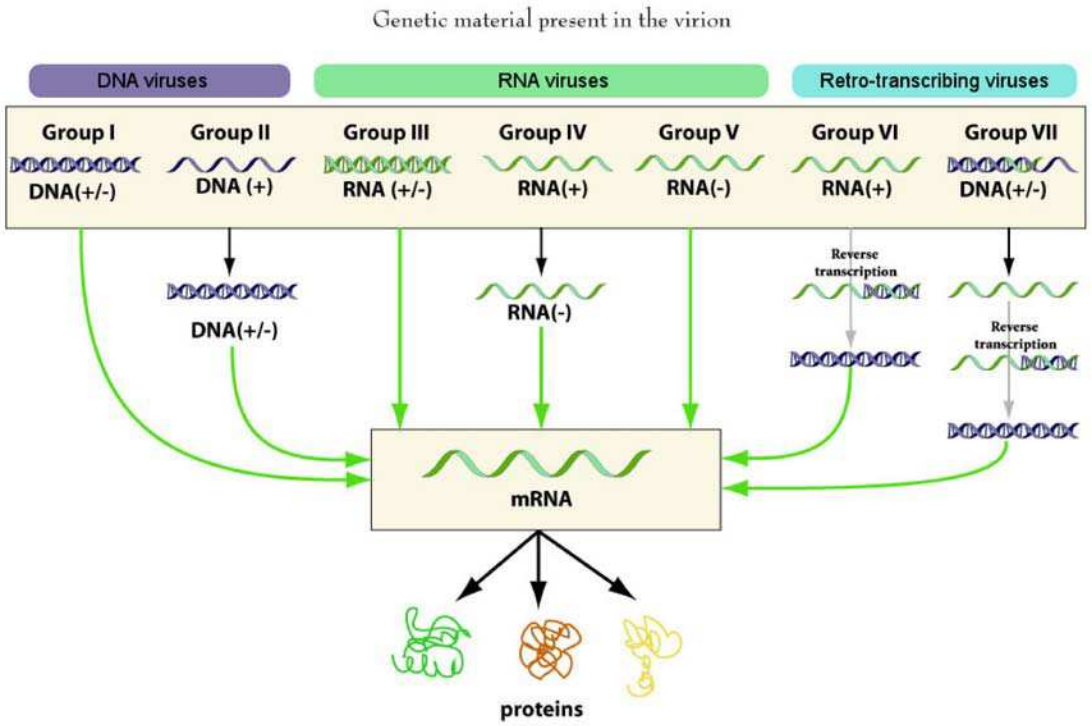


Figure In.1 : Classification de Baltimore basée sur la nature du génome viral et le mécanisme utilisé pour la transcription. *Figure adaptée de Viralzone ©.*

petites des bactéries (et en premier lieu *Mimivirus*), et enfin les liens et similarités frappantes entre les rétrovirus et les transposons décrits par Barbara Mc Clinktock au milieu du siècle.

En dépit de leur nature variée, les virus constituent ainsi des entités biologiques fondamentalement différentes des formes de vie décrites sur Terre et dépendantes du monde cellulaire, dont ils sont des parasites obligatoires. L'existence d'un ensemble de propriétés communes au monde viral a permis l'élaboration d'une liste de caractéristiques largement acceptée :

- la taille des particules virales est très variable, mais inférieure à la taille des cellules
- les virus contiennent un génome sous forme d'acides nucléiques (ADN ou ARN)
- la taille de ces génomes est très variable (de quelques kilobases à plusieurs mégabases), mais généralement inférieure à celle des génomes cellulaires
- les virus ne peuvent se répliquer qu'au sein de la cellule hôte
- leur forme libre, ou "virion", composée d'une capside protéique (parfois entourée d'une enveloppe lipidique) est biologiquement inerte.

2. Diversité du monde viral

Les estimations les plus récentes font état de plusieurs millions d'espèces sur la planète pour les organismes cellulaires (Mora *et al.*, 2011). En considérant que la plupart des formes de vie (si ce n'est la totalité) sont parasitées par au moins un virus, il est vraisemblable qu'un nombre comparable d'espèces virales existe. L'établissement d'une classification de la diversité des virus est toutefois largement contrainte par l'absence de marqueur unificateur du monde viral, comme l'ADN ribosomal pour le monde cellulaire. Les particules virales étant formées d'acides nucléiques encapsidés, deux niveaux principaux de cette diversité ont été étudiés afin d'établir des systématiques virales : la morphologie du virion, et le génome viral.

Diversité morphologique

Les structures des capsides virales sont le plus souvent observées en microscopie électronique ((Marshall, 2012); Figure In.2.A). Cette dernière peut être associée à des techniques comme la microscopie à force atomique qui permet de visualiser la topographie d'une surface (Kuznetsov *et al.*, 2008), la cryofixation (cryomicroscopie, (Veesler *et al.*, 2012)), permettant de mieux préserver la structure native de l'entité observée, et la tomographie qui permet d'obtenir une image en 3 dimensions (Figure In.2.B ; (Ibiricu *et al.*, 2011)). L'autre principale méthode utilisée pour déterminer la structure des capsides virales

est la cristallographie en rayons X, qui permet d'obtenir une image très précise mais requiert une fixation de l'échantillon sous forme de cristal (Marvin Seibert *et al.*, 2011).

Formes et symétries principales des capsides virales

Les capsides des virions sont formées d'un grand nombre de protéines regroupées en sous-unités appelées "capsomères". De manière générale, le nombre de formes de base retrouvées dans le monde viral est relativement restreint (Ackermann, 2007; Abrescia *et al.*, 2012). Les capsides virales sont quasiment toujours symétriques, que ce soit une symétrie icosaédrique, hélicoïdale, ou une combinaison des deux, même si certains virions arborent des formes plus complexes (Figure In.2.A).

Les **capsides de type icosaédrique** sont les plus fréquemment retrouvées (tout type d'hôtes et de matériel génétique), la forme d'icosaèdre (polyèdres à 20 faces triangulaires équilatérales, 30 arêtes et 12 sommets) étant la forme la plus adaptée pour assembler des sous-unités identiques. En effet, l'icosaèdre est le polyèdre régulier convexe qui offre un espace maximal pour le génome viral. Les virions à symétrie icosaédrique sont ensuite distingués en fonction du nombre de sous-unités protéiques contenues dans chacune des unités "de base" formant l'icosaèdre. Il existe ainsi toute une variété de capsides icosaédriques (T2, T3, T4, T5, T7, jusqu'à T27), la plus simple étant de type T1, formée de 60 sous-unités protéiques organisées en 12 pentamères.

L'autre forme majeure retrouvée au sein des capsides virales est la **symétrie hélicoïdale**, présente notamment chez le premier virion dont la structure a été déterminée : le virus de la mosaïque du tabac (Figure In.2.A). Ce type de virion a été décrit principalement pour des virus de plantes (*Potyviridae, Alpha- Beta-* et *Gammaflexviridae*) mais a également été observé pour des virus infectant les bactéries (*Inoviridae*) ou les Archaea (*Lipothrixviridae*, *Rudiviridae*).

Le virion peut aussi consister en un assemblage d'une capside icosaédrique avec une "queue" à symétrie hélicoïdale, comme par exemple pour les virus du groupe des *Caudovirales*. D'autres structures plus originales ont été observées, principalement dans les milieux extrêmes tels que les milieux hypersalins ou les sources d'eau chaude (Bath & Dyall-Smith, 1998; Häring *et al.*, 2005; Porter *et al.*, 2005; Ackermann & Prangishvili, 2012). Ainsi, il s'agit surtout de virus d'*Archaea*, pouvant exhiber des formes de crochet, de citron, de goutte, ou de pyramide (Figure In.2.A). En outre, certains virions peuvent prendre plusieurs formes, et sont alors dit pléomorphes (Mathan *et al.*, 1975). Enfin, une enveloppe lipidique est associée à certains virions, parfois agrémentée de structures glycoprotéiques (Julien *et al.*, 2012; Roine & Bamford, 2012).

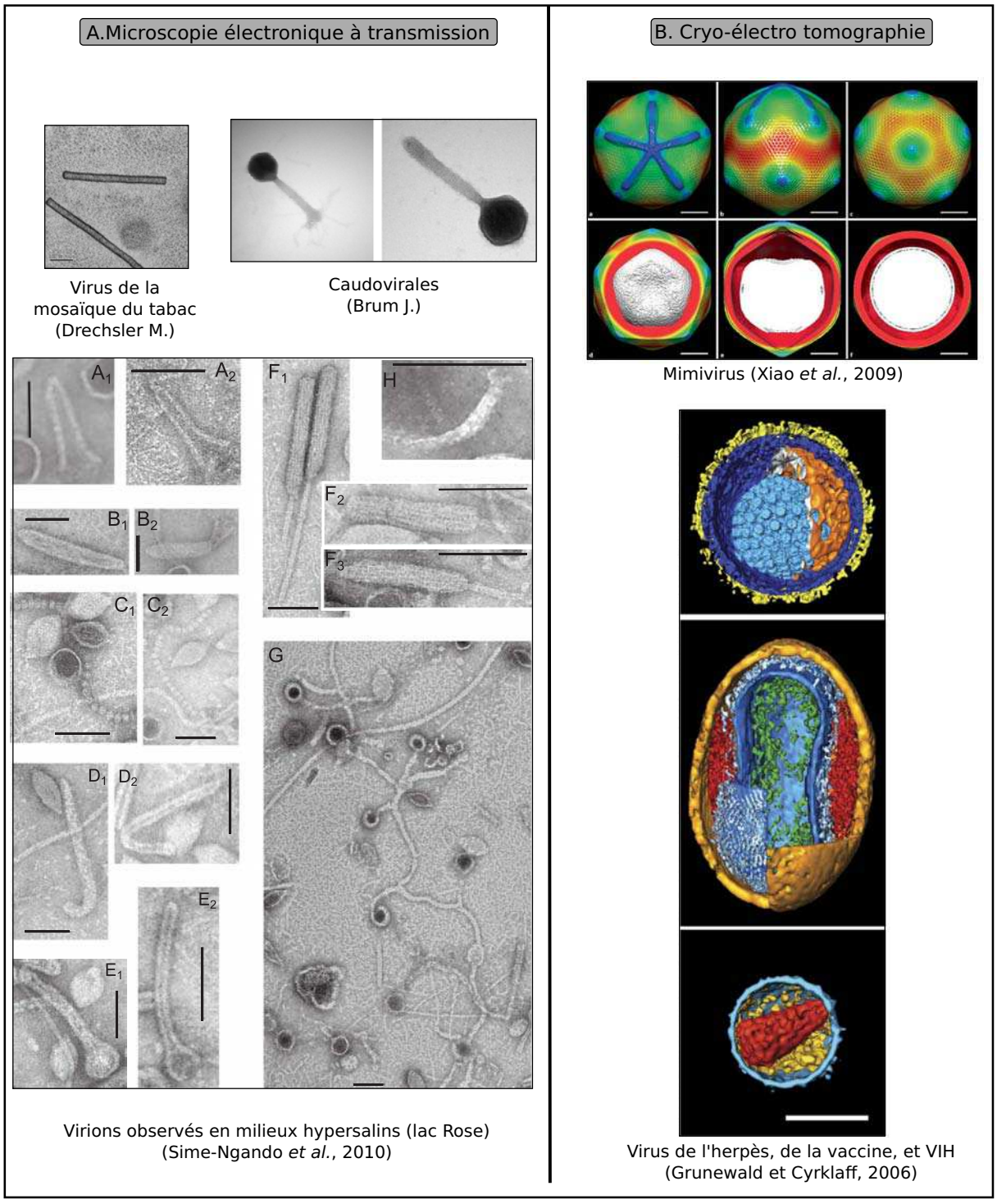


Figure In.2 : A. Virions observés en microscopie électronique à transmission. *Les deux types de symétrie les plus courantes sont présentées en haut : symétrie hélicoïdale (Mosaïque du tabac, Photo M. Dreschler), et icosaédrale (Caudovirales observés au sein de prélèvements océaniques, Photo J. Brum). Le panneau inférieur présente un ensemble de morphologie plus exceptionnelles observées au sein d'un échantillon aquatique hypersalin (A1 et A2 : particules en épingle à cheveux, ; B1 et B2 : particules bacilliformes, C : chaînes de petits globules, D : particules en crochet; E : particules organisées en tige-sphère, F : particules en forme de roseau, G : particules complexes apparemment formé de filaments branchés, et H : structure terminale en crochet observée sur certains filaments ; Image tirée de Sime-Ngando et al., 2010).* ***B. Cryo-électro tomographie de différents virions.*** *Le panneau supérieur présente une vue en surface (A, B, C) et en coupe (D, E, F) du virion de Mimivirus, selon différents angles, et colorés en fonction de la distance au centre de la particule. Le panneau inférieur présente l'organisation interne de trois virions (de haut en bas Herpesvirus, Cowpoxvirus et VIH), colorés de manière à faire apparaître les différentes couches internes. Les barres d'échelle représentent 100 nm.*

Modélisation des capsides virales

Les outils de modélisation de structure des protéines sont également appliqués aux capsides virales, afin d'établir une correspondance entre les séquences de protéines de capside et les structures observées en microscopie. Dans ce cadre, un certain nombre de bases de données dédiées (comme par exemple Viperdb : (Carrillo-Tripp *et al.*, 2009)), ont été créées afin de recenser et comparer les séquences et structures observées des capsides virales, pour dans un deuxième temps pouvoir prédire à partir de nouvelles séquences la capside virale potentiellement formée.

Plusieurs études ont ainsi été réalisées pour mieux appréhender l'évolution des capsides virales (Benson *et al.*, 2004; Bamford *et al.*, 2005; Merckel *et al.*, 2005; Abrescia *et al.*, 2012; Comeau *et al.*, 2012) et différents modèles ont été développés pour comprendre l'assemblage autonome et spontané des protéines en sous-unités puis en virions (Wilber *et al.*, 2009; Johnston *et al.*, 2010; Rapaport, 2010). Ces approches de modélisation de la structure de la capside ont également été utilisées en épidémiologie, par exemple pour prédire les différents variants du virus de la grippe susceptibles d'apparaître (Liao *et al.*, 2008).

Diversité génomique

Au-delà de la diversité morphologique, les techniques d'isolement et de culture de virus ont permis d'accéder à un second niveau de description de la diversité virale avec le séquençage de génomes complets. Les premières analyses génomiques ont été réalisées à partir d'enzymes de restriction, et ont permis, par exemple, de définir en 1973 la carte de restriction du virus simien SV40 (Danna *et al.*, 1973). Peu après, le développement du séquençage de type Sanger a permis d'obtenir les premiers génomes complets.

Séquençage de génomes viraux complets

Les premiers génomes viraux séquencés furent ceux de l'Enterobacteria phage MS2 (Fiers *et al.*, 1976), bactériophage à ARN, puis du phage à ADN simple brin Enterobacteria phage PhiX174 (Sanger *et al.*, 1978), ces deux virus étant cultivés sur *Escherichia coli* et possédant de petits génomes (3 569 pb pour MS2, 5 386 pb pour PhiX174). Après le séquençage du premier virus eucaryote, *Simian virus 40* (Reddy *et al.*, 1978), la gamme de taille de génomes viraux n'a cessé de s'étendre, avec le séquençage de virus géants comme les Herpesvirus, jusqu'à *Megavirus Chiliensis* dont le génome dépasse la mégabase et présente une complexité proche de celle du génome d'une petite bactérie (Arslan *et al.*, 2011).

Ces approches d'isolement, culture, puis séquençage du génome ont apporté des informations essentielles sur les virus et leur diversité génétique. Tout d'abord, une

extraordinaire diversité génétique au sein du monde viral a été mise en évidence : un très grand nombre de séquences ne ressemblant à aucun gène décrit jusqu'à maintenant, chez les virus ou chez les organismes cellulaires, ont été détectées au sein de ces nouveaux génomes. Cette tendance est d'autant plus frappante qu'elle a également été observée entre deux virus isolés à partir du même hôte, et concerne à la fois les éléments fondamentaux des virus et les gènes dits “accessoires”.

Diversité génétique des éléments viraux fondamentaux

Au niveau des **protéines majeures de capside**, il est relativement fréquent de ne retrouver des similarités qu'au niveau des séquences protéiques et de la structure tri-dimensionnelle, sans pouvoir véritablement confirmer ces similarités au niveau de la séquence d'acides aminés en elle-même (et encore moins au niveau de la séquence génomique). De plus, le même type de capside peut être codée par différents types (ou lignées) de gènes. Ainsi, au moins quatre lignées différentes et sans lien évolutif ont été recensées pour les capsides icosaédriques (Abrescia *et al.*, 2012).

De même, les **modules génétiques de réplication** de génomes viraux, s'ils contiennent souvent les même éléments, comprennent une grande diversité pour chaque gène et chaque fonction. Pour les virus à ADN double brin, le complexe de réplication est le plus souvent composé d'une ADN polymérase, d'une ou plusieurs hélicases, d'une primase, et parfois de gènes impliqués dans l'initiation de cette réplication. L'analyse des génomes complets a mis en évidence l'existence pour chacune de ces fonctions de 2 à 6 types de gènes. Ainsi, les polymérases virales retrouvées dans les génomes à ADN double brin peuvent être de type polA, rpolB, ppolB, polC, polY ou primPol, les similarités entre ces différents types de gènes assurant la même fonction étant quasiment inexistantes (Krupovic & Bamford, 2009).

Génomes viraux et gènes accessoires

Au-delà des gènes impliqués dans les fonctions fondamentales du génome viral, les génomes complets séquencés montrent l'existence d'un large répertoire de gènes “accessoires” portés par les génomes viraux. Parmi les gènes identifiés, l'une des découvertes majeures issues du séquençage d'un grand nombre de génomes a été la présence récurrente de gènes impliqués dans le métabolisme cellulaire dans les génomes de phages et de virus d'eucaryotes.

En 2000, Hendrix et collaborateurs ont ainsi noté que les génomes des bactériophages semblent évoluer par acquisition successives de gènes provenant pour partie du génome de leur hôte (Hendrix *et al.*, 2000). Ils proposent ainsi d'associer la terminologie “*moron”* (pour “More DNA”, littéralement “ADN supplémentaire”) à ces gènes, et considérèrent que

l'acquisition de gènes accessoires par transfert horizontal puis la sélection de gènes procurant un avantage constitue un des mécanismes de base d'évolution des génomes de phages.

La même année, la présence de gènes impliqués dans le métabolisme du phosphate fut mise en évidence au sein du génome de Roseobacter phage SIOI (Rohwer *et al.*, 2000), suivie en 2004 par la description de gènes impliqués dans la photosynthèse (photosystème II) dans différents génomes de cyanophages (Lindell *et al.*, 2004). Ces observations joueront un rôle essentiel dans l'acceptation générale de l'idée de la présence de gènes impliqués dans le métabolisme cellulaire au sein des génomes de phages. En effet, plusieurs remarques fondamentales peuvent être issues de cette étude. Tout d'abord, les gènes du photosystème sont retrouvés dans trois génomes complets issus de deux types de phages (*Myoviridae* et *Siphoviridae*), témoignant du fait qu'il ne s'agit pas d'un événement isolé. Ensuite, un ensemble de gènes a été retrouvé au sein des génomes de phages, qu'il s'agisse du photosystème core (psbA), de HLI (*high-lights inducible protein*), de plastocyanines ou de ferredoxines. Enfin, les analyses phylogénétiques ont démontré sans ambiguïté que ces gènes sont d'origine cellulaire, et ont été transférés à de multiples reprises entre le phage et la cyanobactérie.

Différents gènes impliqués dans le métabolisme fondamental de la cellule seront par la suite mis en évidence dans de nouveaux génomes de phages, y compris des gènes impliqués dans le cycle de Calvin et la voie des pentose phosphate (Thompson *et al.*, 2011). Des situations similaires ont été décrites pour les virus d'eucaryotes, avec par exemple l'observation d'une voie de biosynthèse lipidique complète transférée entre le génome d'une algue et un génome viral (Monier *et al.*, 2009). Enfin, l'analyse des génomes de virus géants tels que *Mimivirus* ou *Megavirus* a également révélé la présence de manière récurrente de gènes associés au compartiment cellulaire tels que les ARN de transfert (Legendre *et al.*, 2011).

La majorité des gènes décrits dans les génomes viraux séquencés restent malgré tout non caractérisée à l'heure actuelle, et il est compliqué de prédire leur fonction, voire même de déterminer avec certitude s'il s'agit de véritables gènes. Ces gènes non caractérisés (ou ORFans, pour *orphan ORFs* ou cadres de lecture ouverts orphelins *sensu* (Yin & Fischer, 2008) sont de plus différents entre les génomes viraux complets disponibles : 30 % n'ont aucune similarité avec un gène connu, qu'il soit viral ou cellulaire, et seulement 20 % d'entre eux présentent des similarités avec des gènes de génomes cellulaires (Yin & Fischer, 2008).

Génomique comparative appliquée aux virus

Différentes approches de génomique comparative ont permis de mieux appréhender les mécanismes évolutifs responsables de cette diversité génétique virale. Ce type d'analyse a été réalisé à différentes échelles.

Tout d'abord, ces comparaisons de génomes ont été appliquées à différentes souches du même virus pour mieux décrypter l'évolution fine d'une population. En 2005, Carillo et collaborateurs ont par exemple publié l'analyse de 103 génomes complets de *Foot and mouth diseases virus* issus du monde entier, en observant particulièrement la structure du génome, y compris les régions non codantes (Carrillo *et al.*, 2005). Par la réalisation de phylogénies multiples, ils ont ainsi pu montré qu'il existait une évolution congruente des gènes structuraux, et une perte de signal pour les autres gènes. Ce type d'analyse a également été mené au sein de groupes plus larges, comme par exemple au niveau d'une famille de phages (Cresawn *et al.*, 2011).

D'autres approches ont été développées pour étudier l'évolution des génomes à l'échelle des grands groupes de virus, notamment au niveau des phages bactériens. Ces approches ont tout d'abord visé à organiser les gènes de phages en groupes d'orthologues (ou POGs), selon le principe de regroupement adopté pour le classement des groupes d'orthologues cellulaires (ou COGs ; (Liu *et al.*, 2006)). L'analyse de ces POGs révéla plusieurs tendances majeures de l'évolution des génomes de phages, notamment le faible taux de paralogues et la forte spécificité des gènes de phages au domaine viral (Kristensen *et al.*, 2011, 2013).

Un niveau d'intégration supplémentaire à été proposé par Lima-mendez et collaborateurs, à nouveau dans le cadre d'une analyse globale de l'ensemble des génomes de phages bactériens (Lima-Mendez *et al.*, 2008). Dans un premier temps, les gènes ont été considérés au sein d'un réseau permettant de détecter des groupes (ou "clusters") de gènes homologues, puis les génomes sont considérés à leur tour comme les nœuds d'un réseau, reliés par la présence en leur sein de gènes appartenant à un même groupe. Plusieurs résultats majeurs ont été obtenus par ces auteurs, confirmant pour la plupart les observations réalisées au niveau de groupes plus restreints. Une très forte connectivité a ainsi été notée entre les génomes de phages, la vaste majorité des phages séquencés partageant au moins un gène (Figure In.3). De véritables génomes "chimères", partageant un nombre important de gènes avec deux types de phages n'ayant rien en commun, ont aussi pu être mis en évidence. Enfin, les groupes taxonomiques reconnus et décrits dans les bases de données correspondent bien aux résultats de clusterisation basée sur le réseau de gènes, confirmant que si certains phages distants peuvent partager des gènes, les approches phylogénomiques permettent de reconstituer une histoire évolutive cohérente.

Le même type d'approche a été appliqué à un ensemble plus vaste de séquences incluant les génomes de phages bactériens mais aussi différents prophages détectés au sein de génomes bactériens ainsi que divers plasmides et transposons. Ces séquences ont été réunies au sein de la base de données ACLAME, intégrant les phages au sein d'une population de séquences mobiles ("mobilome" ; (Leplae *et al.*, 2010)). Les résultats des comparaisons de séquences entre plasmides, transposons, et prophages ont démontré que s'il n'existe pas de gène commun à l'ensemble des éléments génétiques mobiles, il existe un certain nombre de modules de gènes conservés au sein d'éléments de nature différente (par exemple entre plasmides et phages). Ces résultats ont ainsi apporté des éléments supplémentaires en faveur de l'hypothèse de liens évolutifs forts entre les virus et les autres structures génétiques mobiles, et ont renforcé l'idée que des transposons ou des plasmides puissent être à l'origine de nouvelles familles virales.

L'accès aux génomes viraux complets a permis d'apporter un nouveau regard sur la virosphère, en permettant d'ajouter aux informations sur la structure et le pouvoir infectieux des virus des informations sur leur diversité génétique, leurs potentialités fonctionnelles et leur histoire évolutive. Cependant, ces analyses sont limitées par la complexité d'isoler des virus en culture, principalement liée à la difficulté de cultiver leurs hôtes (Rappé & Giovannoni, 2003; Case & Boucher, 2011). Si des approches indépendantes de la culture, telles que les approches *single-cell*, ont bien été réalisées sur certaines souches virales (Yoon *et al.*, 2011), ces analyses restent exceptionnelles. Ainsi, au début de l'année 2012, 80 % des génomes complets de phages bactériens provenaient de virus isolés à partir de seulement trois groupes d'hôtes (*Gamma Protéobacteria*, *Actinobacteria* et *Firmicutes*), sur les 61 classes reconnues par le NCBI.

Classification et organisation de la diversité virale

Actuellement, trois principales sources d'information sont considérées pour la classification des virus : leur(s) hôte(s), le contenu de leur génome, et la structure de leur capside, tout en sachant que ces données sont interdépendantes (Figure In.4). Néanmoins, les critères utilisés ont été largement modifiés par le passé, et ils évolueront encore certainement avec l'apport de nouvelles connaissances sur la virosphère.

Paramètres utilisés pour la classification des virus

Dès les années 1940, certains auteurs comme Frederick Bawden proposent un regroupement de certains virus, principalement sur la base de la forme de leur capside

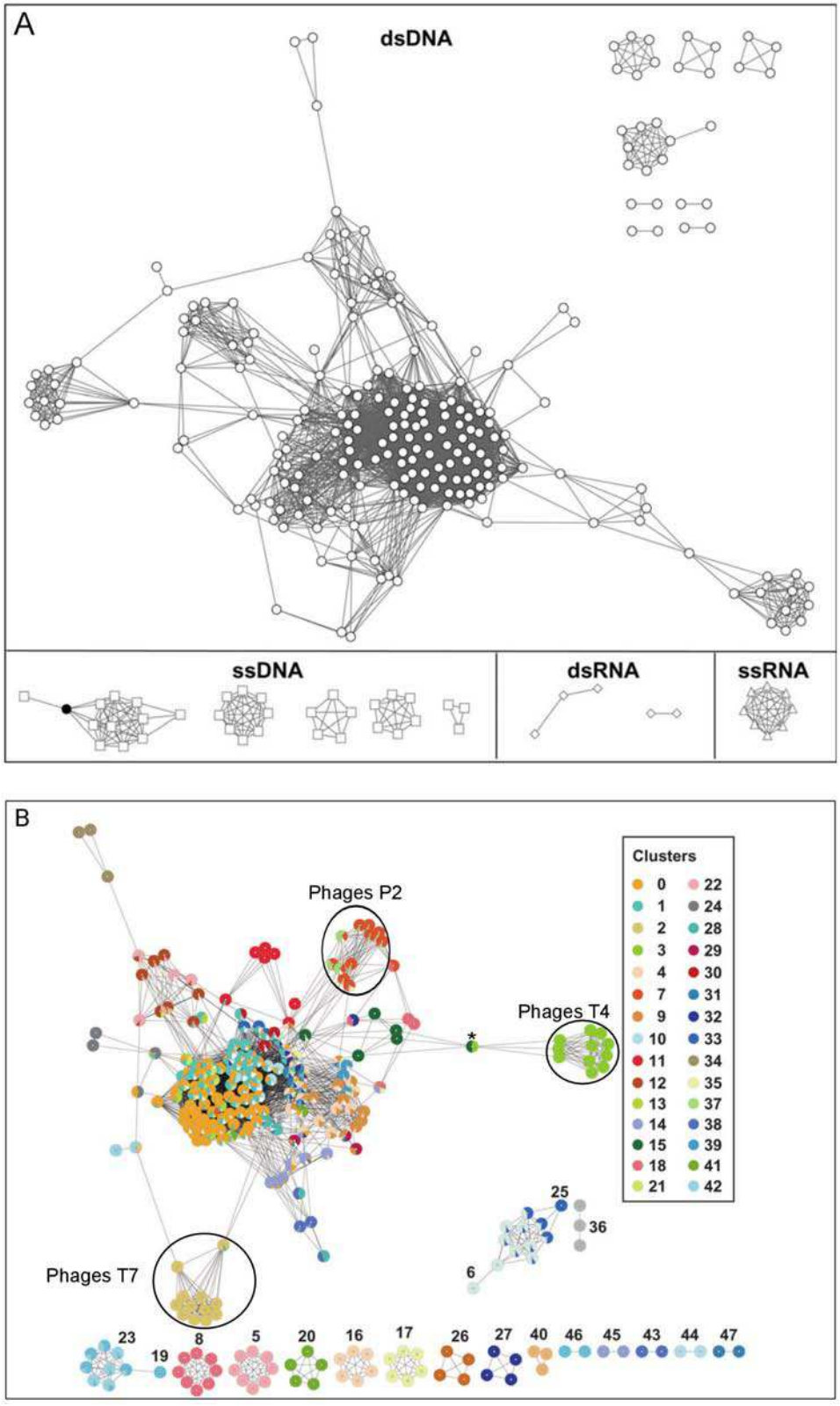


Figure In.3 : Représentation des génomes de phages en réseau sur la base de la composition en gènes (adaptés de Lima-Mendez et al., 2008). *Réseau comprenant l'ensemble des génomes de phages, reliés lorsqu'ils présentent une similarité significative. A : Les phages sont dans un premier temps regroupés par nature de génome, et le type de génome est indiqué pour chaque groupe. B : Pour le même réseau, les nœuds sont colorés en fonction des clusters génomiques auxquels ils sont reliés (basés sur la composition en gènes). Les clusters correspondant à différents groupes de phages connus sont entourés, et le bactériophage T5, qui relie les phages de type T4 (cluster 3) à des pages de type lambda (cluster 15) est indiqué par une astérisque.*

(Bawden, 1941). Les premiers groupes ainsi répertoriés comprennent les *Herpesvirus* (1954), les *Myxovirus* (1955), les *Poxvirus* (1957) et plusieurs groupes de virus filamenteux de plantes (1959). Dans les années 1960, l'explosion du nombre de virus décrits conduisit à la création d'un Comité International pour la Nomenclature des Virus (ICNV), qui fut fondé en 1966 lors du Congrès International de Microbiologie de Moscou (ce comité deviendra le Comité International pour la Taxonomie des Virus ou ICTV en 1973). L'idée directrice retenue pour l'établissement de cette taxonomie universelle des virus est celle d'une classification hiérarchique basée sur une série ordonnée de critères depuis le type d'acide nucléique (correspondant à la classification de Baltimore), puis la symétrie de la capside, la présence (ou absence) d'enveloppe, etc.

L'ICTV publie régulièrement des rapports mettant à jour la taxonomie des virus, et référence aujourd'hui plus de 2 618 espèces au sein de 96 familles. Cette taxonomie se base principalement sur des analyses phylogénétiques (préférentiellement plusieurs phylogénies concordantes de plusieurs gènes) et les grands groupes d'hôtes. Toutefois, une augmentation importante du nombre de familles virales est rapportée, sans qu'il soit possible de les rassembler de manière stricte et rigoureuse au sein de groupes plus larges, et un nombre toujours très important de virus restent non classés (Fauquet & Fargette, 2005).

En 2012, Abrescia et co-auteurs ont proposé une nouvelle classification basée sur la structure de la capside, s'appliquant principalement aux virus à symétrie icosaédrique (Abrescia *et al.*, 2012). Cette classification se base sur le constat qu'un nombre restreint de structures de capsides est observé dans la nature, et part de l'hypothèse que la structure de la capside virale est l'élément le plus fortement contraint d'un virus, de par le nombre limité de possibilités physiques de repliement des protéines. De plus, la plupart des différentes formes de capsides sont retrouvées dans des virus infectant les trois grands domaines de la vie. Ainsi, l'espace déterminé par l'ensemble des structure de capsides semble plus restreint que l'espace décrit par l'ensemble des séquences génomique virale, ce qui devrait à la fois faciliter la classification et permettre l'établissement de liens évolutifs plus ancestraux (Figure In.4).

Limites des classifications actuelles

Dès 1989, l'ICTV a énoncé une définition volontairement générale de l'espèce virale : "*A virus species is a polythetic class of viruses that constitutes a replicating lineage and occupies a particular ecological nich*e". Cette définition reste large et laisse pour mission aux spécialistes de chaque type de virus de définir pour son groupe d'expertise ce que doit être une espèce plus précisément. Par cette décision, l'ICTV reconnaît l'impossibilité d'établir des unités évolutives fondamentales uniformes au sein du monde viral.

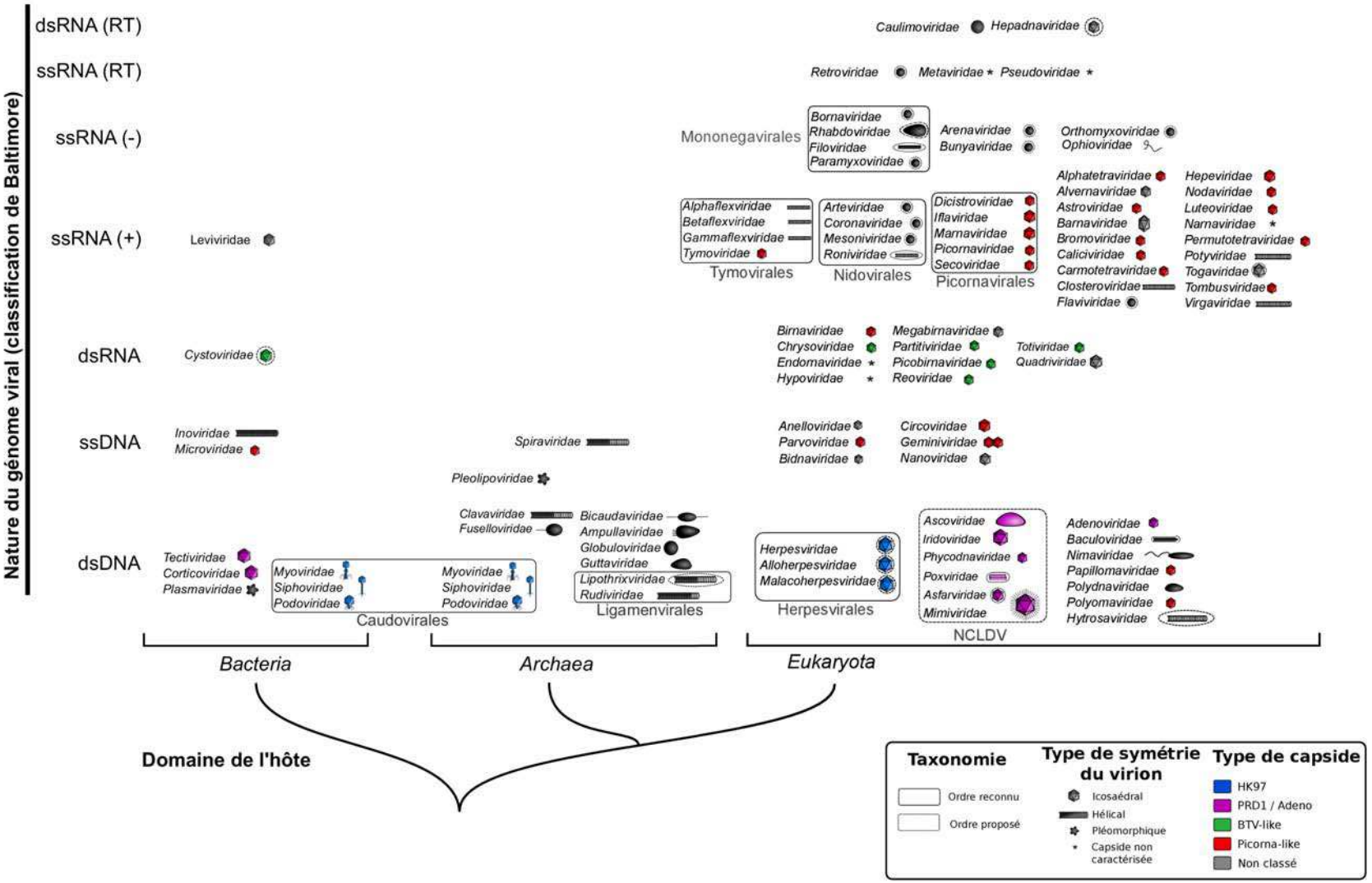


Figure In.4 : Schéma récapitulatif des différentes familles virales. *Un schéma du virion est associé à chaque famille, et différentes classifications des virus sont indiquées : les ordres reconnus (trait plein) ou proposés (trait discontinu) pour l'ICTV, en colonne une classification par le domaine de l'hôte, en ligne par la nature du génome (classification de Baltimore), et enfin une classification par le type de capside sur les schémas de structure de virions (classification issue d'Abrescia et al., 2012). La famille des Pleolipoviridae a été proposée mais non encore approuvée par l'ICTV (nov. 2012).*

L'établissement d'une classification universelle des virus est ainsi une tâche extrêmement compliquée, voire potentiellement impossible, et le croisement des informations de l'ICTV (génomique et hôte) avec les données de structures de capside ne sont que très rarement concordantes. La lignée de capside définie comme “BTV-like” (Figure In.4, capsides vertes) est un exemple de groupe homogène : formée uniquement de virus dont le génome est sous forme d'ARN double brin, elle concerne la grande majorité des virus possédant ce type de génomes. Toutes les familles décrites au sein de ce groupe infectent des eucaryotes (des invertébrés aux humains), et une seule famille infecte des bactéries. Toutefois, cet exemple de concordance entre une nature de génome et une structure de capside reste une exception.

Ainsi, la lignée de capside définie comme “Picorna-like” (Figure In.4, capsides rouges) regroupe des virus à ARN et à ADN, simple et double brin dans les deux cas. Les lignées “HK97” et “PRD1/Adeno” (Figure In.4, capsides bleues et violettes) sont limitées aux virus à ADN double brin, mais englobent encore une fois des virus infectant différents domaines de la vie (bactéries et eucaryotes pour PRD1, bactéries eucaryotes et archées pour HK97). Enfin, les quelques groupes formellement définis et reconnus notamment par l'ICTV ne correspondent pas toujours à ces classifications : si les *Caudovirales* et les *Herpesvirales.* semblent homogènes en matières de nature du génome et de structure de capside, les *Mononegavirales, Nidovirales,* ou *Tymovirales* regroupent des virus à capside icosaédrique et d'autres dont la capside est hélicoïdale.

Les analyses basées sur les génomes complets et les virus isolés ont ainsi révélé certaines caractéristiques fondamentales de l'histoire évolutive des virus et permis la mise en place de classifications globales du monde viral. L'ensemble de ces résultats permet de prendre conscience de la complexité des mécanismes d'évolution en œuvre au sein du monde viral, bien loin de l'image initiale du virus en tant qu'objet génétique simple évoluant uniquement par mutation ponctuelle. Toutefois, la vision du monde viral *via* les virus isolés et les génomes complets séquencés est nécessairement fragmentaire et incomplète. Elle est notamment principalement orientée autour des virus d'intérêt clinique ou agronomique, qui en constituent pourtant une fraction limitée. Ainsi, des pans entiers de la virosphère, comme les phages à ARN et à ADN simple brin sont encore très peu décrits, et le contenu du pangénome viral en termes de diversité génétique et de fonctions codées est encore largement méconnu.

3. Les virus en écologie

Les virus occupent une place importante au sein des écosystèmes et des réseaux trophiques, en particulier ceux infectant les micro-organismes procaryotes et eucaryotes, fortement impliqués dans les cycles biogéochimiques majeurs et composant la majeure partie de la biomasse planétaire. Dans le cadre d'études écologiques, les virus sont ainsi avant tout considérés dans leur rapport avec l'organisme hôte. Il est donc indispensable de décrire tout d'abord les relations et interactions au niveau du couple individuel virus-hôte, avant de pouvoir appréhender les différents impacts des virus au niveau des communautés et des écosystèmes.

Complexité des interactions virus – hôte

Cycle lytique

Le cycle de vie ou de multiplication d'un virus peut être extrêmement variable en fonction du type de virus, de l'hôte, ou des conditions environnementales. Toutefois, le cycle principalement observé, dit cycle lytique, peut être décrit en six étapes principales retrouvées de façon quasiment universelle. Le virus doit tout d'abord s'attacher à la membrane de son hôte, puis pénétrer la cellule. Le génome doit ensuite être libéré au sein de la cellule (dans certains cas, notamment les *Caudovirales,* le génome est libéré au sein de l'espace intracellulaire sans que le virion ne pénètre la cellule, mais les mécanismes de cette "injection" sont encore débattus, (Grayson & Molineux, 2007)).

Une fois le génome disponible au sein de la cellule, une étape de biosynthèse des composants viraux (à la fois protéique et nucléique, principalement pour former la capside d'un côté et répliquer le génome viral) prendra place. Les capsides seront ensuite assemblées, et le génome viral encapsidé, avant un relargage des virions dans l'espace extracellulaire. Ce modèle résume ainsi les caractéristiques principales et indispensables d'un virus, à savoir la possibilité de générer des virions et d'encapsider son génome, ainsi que la nécessité d'utiliser au moins partiellement la machinerie cellulaire de l'hôte pour y parvenir.

L'effet immédiat du virus sur son hôte consiste donc en une mortalité induite par l'infection virale lors des cycles lytiques, *i.e.* menant à la lyse de la cellule infectée. Cet effet néfaste des virus entraîne le développement de systèmes de défense et de contre-défense entre l'hôte et le virus, dans ce qui est souvent désigné comme une "course à l'armement" (Stern & Sorek, 2011). Si ces mécanismes sont chez les eucaryotes pluricellulaires des éléments participant au système immunitaire, de tels systèmes en "miroirs" entre virus et hôtes ont été décrits chez tous les organismes, y compris les bactéries et archées (notamment les systèmes

CRISPR / Cas, (Terns & Terns, 2011)). Toutefois, l'impact d'un virus sur son hôte ne se limite pas à une atteinte du développement ou un effet sur la santé de l'organisme infecté, même s'il s'agit du premier élément de découverte des virus.

Cycles lysogéniques et chroniques

A l'opposé du cycle lytique se trouve le cycle lysogénique, qui implique une phase d'intégration du génome viral au sein du génome de l'hôte. Ce cycle est principalement rencontré et décrit pour les bactériophages, même si certains virus eucaryotes peuvent aussi s'intégrer au génome de leur hôte (Lefeuvre *et al.*, 2011). Le génome de phage intégré est nommé prophage et peut suivre deux destinées : soit il sera libéré à nouveau en tant que génome complet (avec une composition en gène identique ou quasi identique), soit il va perdre peu à peu des gènes, subir des insertions de gènes bactériens, et ne pourra plus alors former de virions (on parle alors de prophages “déficients”).

Ce type de cycle semble très fréquent, puisque les prophages (complets ou non) sont retrouvés en abondance dans les génomes bactériens entièrement séquencés, et peuvent représenter jusqu'à 20 % du génome de l'hôte (Canchaya *et al.*, 2004). La plupart de ces prophages identifiés semblent toutefois être déficients, et ne peuvent probablement pas retourner à l'état de virus libre. Ils peuvent néanmoins toujours jouer un rôle important en apportant de nouveaux gènes à l'hôte.

Certains cycles viraux vont au-delà de la lysogénie et permettent une production de virions infectieux sans destruction de la cellule hôte, dans ce qui est alors appelé un cycle chronique. Ce type d'interaction a principalement été décrit pour des virus humains comme le virus de l’herpès (Knipe & Cliffe, 2008), mais a récemment été mis en évidence au sein d'eucaryotes unicellulaires. Ainsi, il a été observé chez certaines algues l'existence d'un état de coexistence entre l'hôte et le virus, dans lequel la croissance de l'hôte est maintenue en parallèle d'une production réduite mais existante de virus, sans que le génome de ce dernier ne soit intégré au génome cellulaire (Thomas *et al.*, 2011).

Effets bénéfiques des infections virales

Les effets potentiels des virus sur le métabolisme et la survie de leurs hôtes sont donc bien plus complexes que la simple destruction de l'hôte par le virus, et les gènes apportés par le génome viral peuvent même procurer un véritable avantage à la cellule hôte. Dans le cas de la bactérie *Vibrio Cholera* par exemple, les souches dites virulentes (capables de produire la toxine du choléra) ont vraisemblablement acquis le(s) gène(s) codant pour cette toxine par transfert horizontal à partir d'un génome de phage, le Vibrio Phage CTXφ de la famille des *Inoviridae* (Faruque & Mekalanos, 2003). De manière plus générale, les prophages semblent

impliqués dans plusieurs cas d'acquisition de pathogénicité chez des bactéries (Busby *et al.*, 2013). Chez les eucaryotes, il a été montré qu'un champignon parasite de plante isolé à partir des sources d'eau chaude de Yellowstone n'était capable de résister à ces températures extrêmes qu'en présence d'un virus spécifique (Márquez *et al.*, 2007). La présence du virus permet ainsi de conférer non seulement au champignon mais aussi à son hôte végétal cette capacité de résistance à la chaleur.

De manière plus générale, la valeur sélective (ou *fitness*) de l'hôte peut être améliorée par la présence d'un virus. Dès 1977, Lin, Bitner et Edlin notèrent que les cultures d'*Escherichia coli* qui contiennent le bactériophage lambda sous forme lysogénique présentent un meilleur taux de reproduction, du à un métabolisme plus actif de ces bactéries (Lin *et al.*, 1977). Plus récemment, la découverte de gènes impliqués dans différents métabolismes cellulaires fondamentaux au sein des génomes de phages a fait émerger l'idée d'une optimisation par le phage du métabolisme de la cellule infectée (Thompson *et al.*, 2011). Différentes études (principalement chez les cyanophages) ont démontré que ces gènes du métabolisme et notamment les protéines principales du photosystème étaient effectivement exprimées durant l'infection, et remplaçaient les exemplaires cellulaires de ces même protéines, dont le nombre décline au fur et à mesure de l'infection (Lindell *et al.*, 2005).

Les interactions virus-hôtes sont donc très complexes, et peuvent prendre des formes extrêmement différentes. Ainsi, les forts taux de mortalité dans la nature semblent être observés lors des cas d'associations nouvelles entre un virus et son hôte, et il semble qu'à terme, un équilibre s'installe et qu'une coexistence se mette en place entre les virus et les organismes hôtes (Thyrhaug *et al.*, 2003; Ebert & Bull, 2007). Généralisées au niveau des communautés microbiennes, ces interactions peuvent influencer le devenir des différents organismes et moduler la structure et la diversité des communautés. L'étude de ces interactions doit de plus prendre en compte les paramètres environnementaux : certaines études postulent ainsi qu'un équilibre existe entre les cycles lysogéniques et lytiques pour les phages aquatiques, dépendant notamment de la taille des populations d'hôtes et donc indirectement du statut trophique de l'environnement (Payet & Suttle, 2013). L'étude des communautés virales en tant que compartiment écologique est donc primordiale puisqu'elles vont profondément influencer le fonctionnement des écosystèmes.

Impact des virus a l'échelle des communautés

Différentes approches ont été développées afin d'étudier les communautés virales environnementales dans leur ensemble et dépasser l'étude de modèles virus-hôtes isolés. Elles

s'appuient généralement sur les caractéristiques structurales des capsides virales, afin de les isoler à partir d'un échantillon naturel complexe *via* une combinaison d'étapes de filtration, précipitation et concentration.

L'analyse des communautés virales environnementales a dans un premier temps été réalisée par l'intermédiaire d'observations en microscopie électronique à transmission (MET) associées à des comptages en microscopie à épifluorescence ou cytométrie en flux. Par la suite, la caractérisation génétique de ces communautés a été rendue possible par des approches d'empreinte génétique, *via* l'utilisation d'électrophorèses sur gel en gradient dénaturant (PFGE, DGGE ; (Sandaa *et al.*, 2010)), ou d'amplification aléatoire (RAPD-PCR, (Helton & Wommack, 2009; Winter *et al.*, 2013)). Enfin, l'étude de gènes marqueurs par PCR a également permis d'obtenir différentes informations sur la diversité de ces gènes, et par extension des organismes qui les abritent, en permettant par exemple le suivi de populations virales environnementales (Short *et al.*, 2011).

Observation de capsides virales au sein d'échantillons environnementaux

L'étude des virus en tant que communautés écologiques dans l'environnement a connu un regain d'intérêt au début des années 1990 lorsque des analyses de comptage ont révélé une présence extrêmement importante de capsides virales dans les échantillons environnementaux, notamment aquatiques (Bergh *et al.*, 1989). A titre d'exemple, des concentrations de 10^8 particules virales par millilitre ont été mesurées dans différents prélèvements océaniques, et on estime que les océans dans leur globalité abriteraient 4.10^{30} virus, soit l'équivalent en carbone d'environ 75 millions de baleines (Suttle, 2005). Si les mesures de ces fortes abondances pourraient être biaisées par certaines limitations méthodologiques (Forterre *et al.*, 2013), la présence importante de virus a été confirmée par l'observation fréquente de cellules infectées par des virus (Suttle, 1994; Weinbauer *et al.*, 2003).

Par la suite, ces observations de particules virales en nombre important ont été multipliées et validées dans tout un ensemble de types d'échantillons, comme les milieux d'eau douce (lacs et rivières), les milieux extrêmes (qu'il s'agisse de fortes salinités, fortes pressions, ou températures particulièrement élevées ou basses), mais aussi dans l'ensemble des organismes vivants, et au sein de ces organismes, dans tous les organes et compartiments. Des quantités importantes de particules virales ont notamment été observées dans les sédiments marins avec une concentration encore plus importante que dans la colonne d'eau, et où les virus gardent leur pouvoir infectieux (Suttle, 2005). Ainsi, les virus sont retrouvés dans l'ensemble des types d'échantillons connus, que ce soit lié à la présence de leur hôte dans le même échantillon ou de manière plus transitoire.

Modèles de distribution des communautés virales

Une très forte diversité a été rapportée au niveau des communautés virales dans l'ensemble des milieux étudiés, notamment aquatiques (Breitbart & Rohwer, 2005). Ces observations suggèrent soit une richesse globale exceptionnelle des virus sur la planète, soit une grande capacité de dispersion des virus, auquel cas ces fortes richesses locales seraient associées à une richesse globale limitée.

Cette capacité de dispersion des virus a notamment été illustrée par la détection de séquences quasi-identiques de phages bactériens au sein de prélèvements espacés de plusieurs milliers de kilomètres, issus de différents types d'environnement (Breitbart *et al.*, 2004b). Toutefois, d'autres études ont à l'inverse mis en lumière une influence des paramètres physico-chimiques, avec une distribution au sein de clades environnementaux (lacustres et marins) observée pour des cyanophages (Short & Suttle, 2005; Wilhelm *et al.*, 2006) et des virus d'algues (Clasen & Suttle, 2009; Short & Short, 2009). Ces répartitions pourraient refléter la structure des communautés d'hôtes, et témoigner d'un effet indirect des paramètres environnementaux sur la distribution des virus (Logares *et al.*, 2009). Ainsi, les facteurs structurant les communautés virales et les liens entre les communautés des différents milieux sont encore mal définis.

Différents modèles ont été proposés pour expliquer la forte diversité locale des communautés virales, notamment celui dit "*seed-bank*" (Breitbart & Rohwer, 2005; Vega Thurber, 2009). Dans ce modèle, un nombre limité de virus est effectivement actif et présent en grande quantité, tandis que le reste de la communauté virale est composé de virus inactifs, chacun présent en quantité limitée. Ce modèle éclaire notamment sous un nouveau jour les concentrations importantes de virus détectées dans les sédiments marins.

Régulation virale des communautés d'hôtes

L'effet le plus direct des virus sur les communautés microbiennes est la régulation des communautés d'hôtes *via* la lyse virale. Dans les milieux océaniques, où un taux élevé de mortalité bactérienne a été observée, l'observation de fortes densités de particules virales a rapidement suggéré l'idée que les communautés microbiennes océaniques étaient très largement infectées par des virus qui, pour le maintien de leur propre communauté, induisaient cette forte mortalité microbienne. Plusieurs modèles ont ainsi été proposés pour expliquer plus en détail ces interactions entre communautés virales et microbiennes, et notamment l'influence des virus dans la régulation et la diversité des communautés d'hôtes.

L'un des plus connus est désigné "*kill-the-winner*", et se base sur un modèle de Lokta-Voltera pour identifier des cycles de développement d'hôte, suivi par le développement du virus associé qui entraîne le déclin de cet hôte (Thingstad, 2000). Dès qu'une population

microbienne domine une niche écologique particulière, des virus spécifiques de cette population limiteront son développement, aboutissant à un maintien de la diversité au sein des communautés de micro-organismes.

Les efflorescences (ou blooms) de phytoplancton illustrent bien ce modèle, en particulier celles d'*Emiliania Huxleyi*, espèce formant des blooms océaniques parmi les plus importants au monde. La disparition parfois très rapide de ces blooms (de l'ordre de quelques dizaines d'heures), a rapidement été imputée aux communautés virales. Cette hypothèse a été confirmée par l'observation d'un nombre important de particules virales, dont certaines ont été identifiées comme de nouveaux virus, au niveau des blooms en déclin (Yoon *et al.*, 2011).

Spécificité d'hôte des virus de l'environnement

La spécificité d'hôte des virus vis-à-vis des organismes microbiens reste aujourd'hui compliquée à appréhender au niveau de la communauté. La plupart des études semblent montrer une grande versatilité des virus, avec certaines souches capables d'infecter un large panel d'hôte, quand d'autres sont spécifiquement associées à un type ou une espèce particulière (Sano *et al.*, 2004; Atanasova *et al.*, 2012; Clerissi *et al.*, 2012).

Les rares études à large échelle semblent tout de même montrer qu'il existe une communauté virale distincte associée à chaque groupe microbien, et que les différences de spécificités interviennent ensuite au sein de ces communautés associées (Weitz *et al.*, 2012). Ce modèle d'associations hôtes-virus imbriqués (ou "*nested*") serait lié à la coévolution entre les communautés d'hôtes et de virus, et à l'apparition successive de résistances et de contre-mesures, dans une dynamique temporelle proche du modèle de la reine rouge ("*red queen hypothesis*").

Coévolution des génomes

Ces interactions entre communautés de virus et d'hôtes et les dynamiques de résistance associées ont fortement influencé l'évolution des génomes de micro-organismes. En effet, le développement continu de ces systèmes de résistance ("course à l'armement") conduit à une adaptation et à une modification des génomes de l'hôte et du virus (Stern & Sorek, 2011). Au niveau des communautés, plusieurs expériences de laboratoire ont montré que ces coévolutions de virus et micro-organismes provoquaient une augmentation de la vitesse d'évolution du génome de l'hôte (Pal *et al.*, 2007), et favorisaient la création de nouveaux gènes et potentiellement de nouvelles fonctions au sein du génome de l'hôte (Pal *et al.*, 2007; Stern & Sorek, 2011; Meyer *et al.*, 2012).

En extrapolant ces résultats aux milieux naturels, et en tenant compte du nombre très important d'infections virales intervenant chaque seconde, il est possible de considérer que les

virus ont fortement influencé l'histoire et l'évolution des génomes des organismes cellulaires. Les interactions virus-hôtes pourraient notamment expliquer les différences parfois importantes observées entre les génomes au sein d'une même espèce (Rodriguez-Valera *et al.*, 2009). Les virus (et en particulier les phages) seraient alors des agents de maintien de la diversité génomique à l'échelle de populations spécifiques, dont l'effet serait ajouté à celui de la sélection par l'environnement (Figure In.5 ; (Rodriguez-Valera *et al.*, 2009)). L'influence des virus serait ainsi visible à deux échelles différentes : au niveau des communautés (modèle "*kill-the-winner*") et au sein de chaque population.

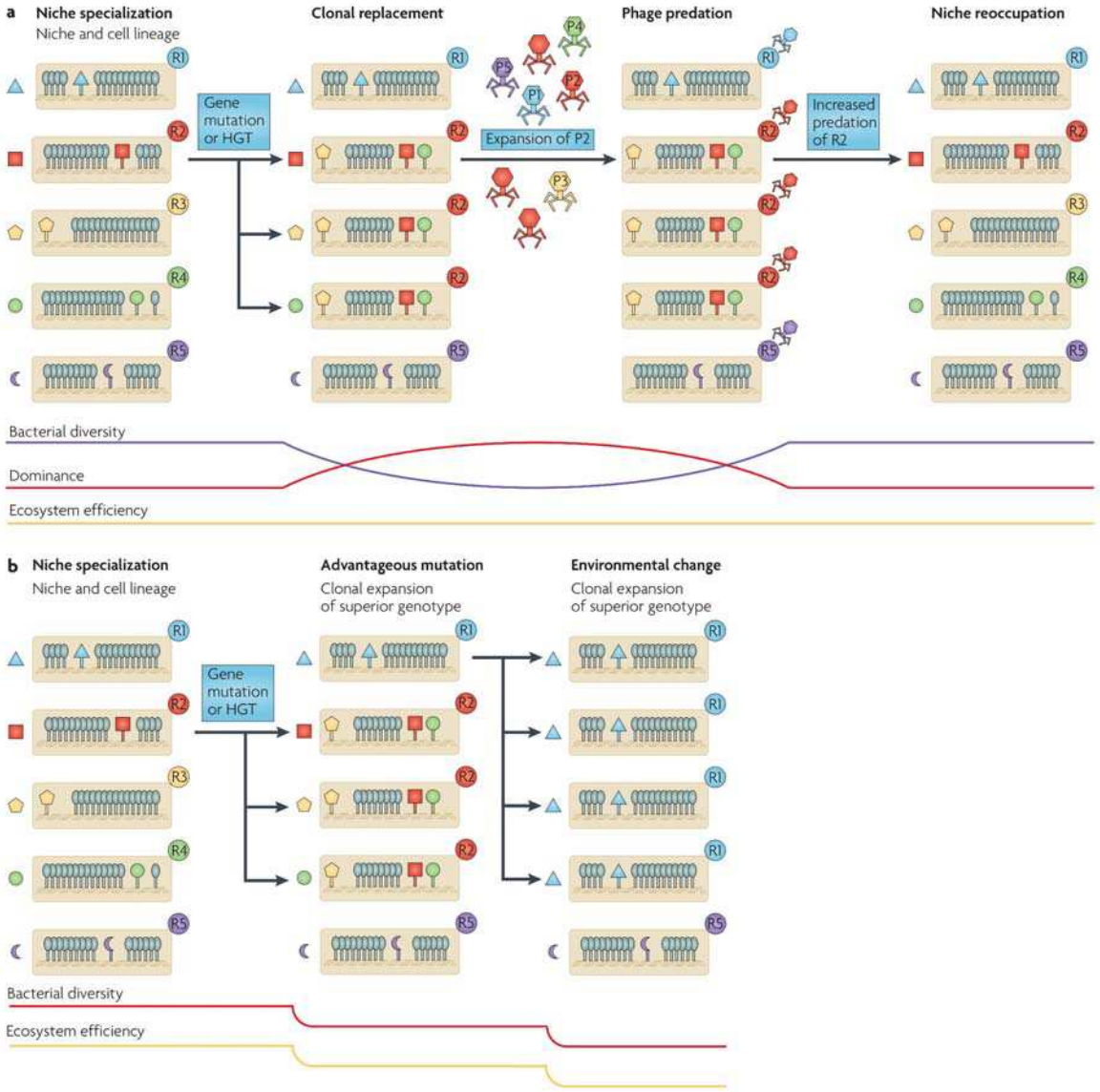


Figure In.5 : Maintien de la diversité bactérienne par la prédation virale (figure issue de Rodriguez-Valera* et al., *2009). *A. Selon le modèle théorique "kill-the-winner", toute souche bactérienne qui, au sein d'une niche écologique où cohabitent plusieurs types bactériens, en viendrait à dominer cet environnement serait alors soumise à une plus forte prédation virale. Au final, l'équilibre entre les différentes bactéries initialement présentes serait ainsi rétabli. B. En l'absence de virus, toute souche présentant une meilleure adaptation va rapidement dominer la niche écologique, réduisant ainsi la diversité bactérienne.*

L'impact des virus sur les communautés microbiennes dépasse donc largement le cadre “simple” de l'infection virale, et touche potentiellement tous les compartiments de la vie cellulaire, et donc le fonctionnement des écosystèmes.

Influence indirecte des virus sur les cycles biogéochimiques majeurs

Du point de vue biogéochimique, il est établi que les virus ont un impact important sur les cycles majeurs (notamment les cycles du carbone et de l'azote), mais aussi sur les flux de nutriments, et participent ainsi aux équilibres globaux de la planète (Fuhrman, 1999). Cet impact des virus est principalement lié aux spécificités de la lyse virale. En effet, la prédation conduit à un transfert des éléments fondamentaux (C, N, P) vers les compartiments supérieurs du réseau trophique alors que les virus vont lyser les cellules infectées, et induire un relargage de ces éléments sous forme de matière organique dissoute. Ces matières organiques dissoutes vont alors être consommées par les micro-organismes, formant ainsi une boucle de recyclage des nutriments, sans passage au compartiment supérieur. La mise en évidence de cette déviation des flux de nutriments due aux virus a mené à l'établissement du concept de “*viral shunt*”.

L’activité virale pourrait à plus grande échelle permettre un maintien des nutriments dans les zones de surface, notamment au sein des océans. En effet, des éléments importants tels que l'azote, le phosphate ou le fer sont plus concentrés au sein des bactéries, qui ont moins tendance à sédimenter que les organismes plus grands comme les eucaryotes unicellulaires. Ainsi, en limitant le transfert de nutriments vers ces eucaryotes qui emporteraient ces éléments essentiels vers les zones profondes et les sédiments, les virus participeraient au maintien de conditions favorables pour les micro-organismes dans les zones de surface océanique (Suttle, 2007).

Les virus constituent ainsi un compartiment écologique complet dont les impacts sur le fonctionnement des écosystèmes et sur les autres membres de la biosphère sont multiples. Il est ainsi primordial de mieux comprendre les facteurs influençant la composition et la distribution de ces communautés virales dans les environnements naturels, et notamment au sein des écosystèmes aquatiques de par leur place centrale dans les cycles biogéochimiques majeurs.

4. La métagénomique appliquée à l'étude de la diversité virale

Les approches métagénomiques sont apparues à la fin des années 1980, bien que le concept de métagénome et le terme en lui-même ne seront formellement définis qu'en 1998 (Handelsman *et al.*, 1998). Cette méthodologie, d'abord utilisée lors de l'étude des communautés microbiennes environnementales, consiste en un séquençage massif et aléatoire des fragments génomiques issus d'un prélèvement. Dans le contexte d'un monde viral fortement diversifié et majoritairement inconnu, la métagénomique apparaît ainsi comme particulièrement pertinente. En effet, elle s'affranchit des limites de la mise en culture, et ne nécessite pas de connaissances préalables sur les séquences étudiées. Ce type d'approche permet à la fois de traiter les communautés virales dans leur ensemble, ce qui est impossible par approche PCR en l'absence de gène marqueur universel pour les virus, et d'accéder aux séquences de génomes viraux, ce que ne permettent pas les approches d'analyse de profil. Ainsi, les séquences métagénomiques constituent une source importante d'informations nouvelles sur la composition et la diversité d'une communauté virale, au niveau taxonomique et fonctionnel.

Il est possible d'étudier la diversité virale à partir de métagénomes visant les micro-organismes dans leur ensemble. Il s'agit alors de détecter et d'analyser au sein de métagénomes ciblant des organismes cellulaires des séquences typiques de génomes viraux séquencés, issues soit de génomes viraux présents dans les cellules étudiées (intégrés ou non au génome cellulaire), soit de virus géants présents dans la fraction de taille microbienne (Monier *et al.*, 2008; Williamson *et al.*, 2008b; Sharon *et al.*, 2009; Comeau *et al.*, 2010; Sorokin *et al.*, 2010). Ce type d'approche présente toutefois deux limites principales : elles ne concernent qu'une fraction très minoritaire des jeux de données étudiés, et permettent uniquement d'analyser des séquences similaires à des virus déjà connus. Il est toutefois possible de générer des métagénomes viraux (ou viromes) donnant accès à l'ensemble de la communauté de virus, en ciblant uniquement les capsides virales et le matériel génétique encapsidé au sein d'un échantillon.

Méthodes de préparation et d'analyse des viromes

Extraction, amplification et séquençage des génomes encapsidés

Afin de décrire une communauté virale dans son ensemble, il est donc nécessaire d'isoler préalablement les virions des micro-organismes. Pour ce faire, une série d'étapes de filtration, précipitation et concentration est généralement réalisée (Figure In .6). Une fois les virions isolés, le matériel génétique encapsidé est libéré (le plus souvent *via* un choc

thermique) et extrait. C'est lors de cette étape qu'un choix sera le plus souvent fait entre l'étude de la fraction ARN et ADN. Malgré les efforts de concentration et les volumes initiaux prélevés souvent importants, il est très rare que la quantité de matériel génétique obtenue soit suffisante pour procéder directement à un séquençage. Ainsi, pour la plupart des viromes, une étape d'amplification généralisée est réalisée.

L'approche la plus couramment utilisée est l'amplification par déplacements multiples de brin (MDA) *via* la polymerase phi29, amplification basée sur un kit d'amorces dites "universelles", et permettant ainsi d'amplifier l'ensemble du matériel génétique isolé à partir de l'échantillon. D'autres approches d'amplification existent mais sont moins générales, comme l'amplification par cercle roulant qui cible les séquences d'ADN circulaire, ou l'amplification par ligation d'une séquence d'adaptateur (ou "*linker*") qui ne concernera que les séquences à ADN double brins. Ces différentes techniques de concentration, de purification et d'amplification sont choisies en fonction du type d'échantillon, de la quantité de matériel disponible, et de la fraction virale d'intérêt. Toutefois, ces techniques ne sont pas

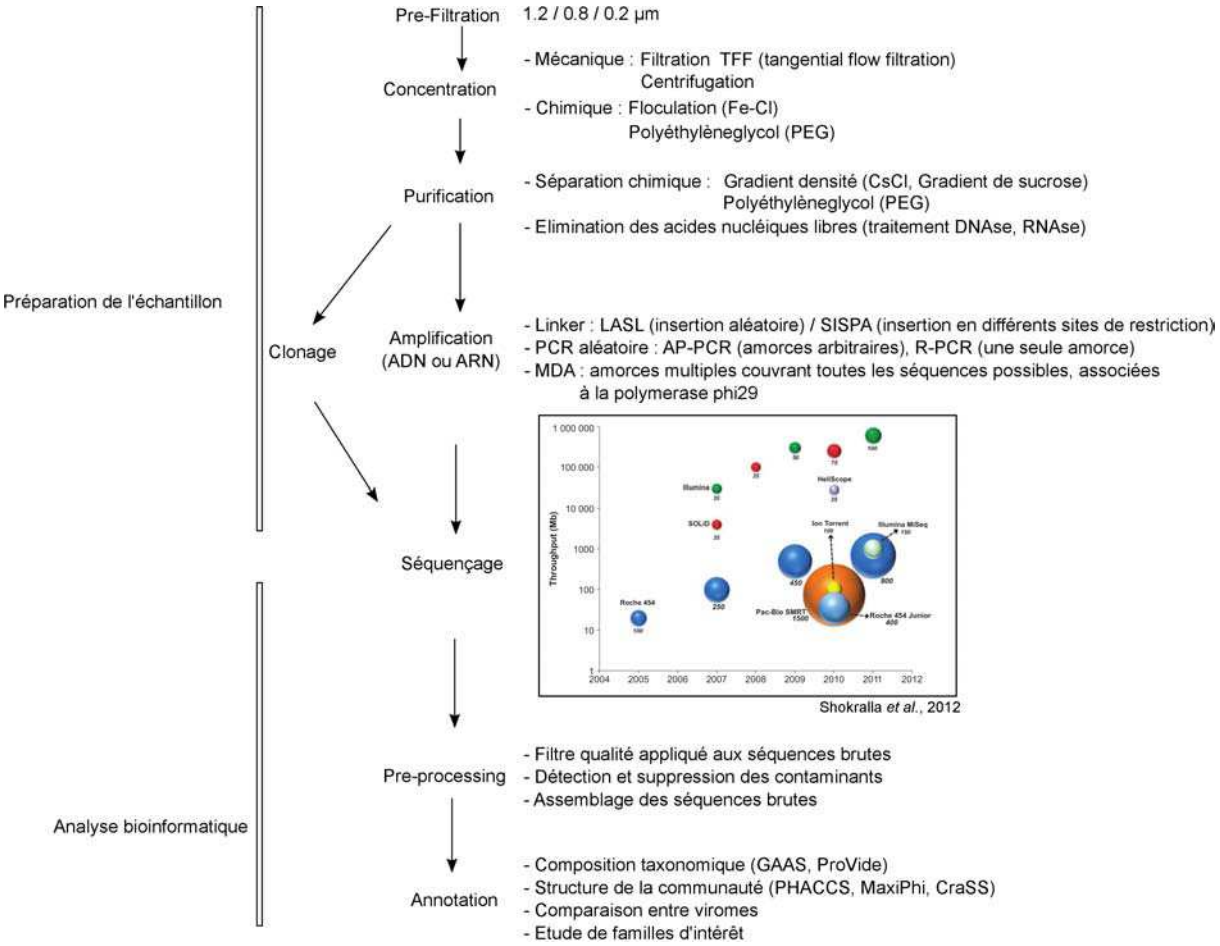


Figure In.6 : Schéma des principales méthodes utilisées pour la préparation des échantillons et l'analyse de séquences de viromes. *Seuls les outils bioinformatiques spécifiquement adaptés aux métagénomes viraux sont indiqués. Les techniques de séquençage à haut-débit majoritairement utilisées pour les métagénomes viraux sont le séquençage 454 (en bleu) et Illumina (en rouge). La figure récapitulant les capacités des différents types de séquençage est issue de Shokralla et al., 2012.*

toutes équivalentes, et peuvent conduire à des différences dans les jeux de données obtenus (Duhaime & Sullivan, 2012; Hurwitz *et al.*, 2012; Solonenko *et al.*, 2013).

Différentes techniques peuvent ensuite être utilisées pour séquencer le matériel génétique isolé. Ces techniques de séquençage ont connu une série d'évolutions très rapide ces dix dernières années, avec l'apparition et le développement des techniques de séquençage à haut-débit (High-Throughput Sequencing, ou HTS). Les HTS ont profondément modifié les masses de données à analyser, en permettant d'obtenir une profondeur de séquençage sans précédent (des centaines de milliers de séquence dans un premier temps, jusqu'à plusieurs centaines de millions aujourd'hui). Contrairement aux étapes de concentration et d'amplification, l'utilisation de différentes techniques de séquençage ne semble pas introduire de biais quant aux résultats obtenus lors de l'analyse de viromes (Solonenko *et al.*, 2013).

Analyse bioinformatique de viromes

Les premiers traitements bioinformatiques d'un virome sont généralement constitués de différentes étapes de filtres et de contrôles de la qualité des séquences, et éventuellement d'un assemblage des séquences brutes en séquences assemblées (ou contigs). Ces différents traitements visent à optimiser les étapes suivantes de l'analyse en limitant la redondance et en générant des séquences les plus longues possibles.

Les analyses les plus couramment réalisées à partir des séquences de viromes peuvent être séparées en quatre grandes étapes : l'établissement de la composition taxonomique de la communauté étudiée, l'estimation de la structure de cette communauté, la comparaison avec d'autres viromes, et enfin l'étude plus spécifique de groupes d'intérêt au sein du virome.

La plupart des outils bioinformatiques utilisés ne sont pas spécifiques des métagénomes viraux, mais plus généralement appliqués lors des études de métagénomes microbiens. Deux logiciels, GAAS (Angly *et al.*, 2009) et Provide (Ghosh *et al.*, 2011), proposent toutefois des algorithmes spécifiquement adaptés aux viromes. Dans les deux cas, il s'agit de déterminer le plus précisément possible la composition taxonomique de la communauté virale étudiée, en prenant en compte respectivement les différences de taille de génomes entre deux virus, et les différentes vitesses d'évolution des gènes utilisés comme marqueurs. Un troisième outil spécifique (PHACCS (Angly *et al.*, 2005)) est dédié à l'estimation de la structure de la population virale à partir de l'analyse de viromes.

Il existe ainsi très peu d'outils spécifiquement adaptés aux données de métagénomique virale, sans doute parce que cette discipline est encore jeune et en développement. En particulier, aucun logiciel ne permet de comparer différents viromes de manière rigoureuse, ou d'étudier la diversité des communautés virales par l'intermédiaire de gènes marqueurs spécifiques.

Apports de la métagénomique pour l'étude des communautés virales

Ces approches de métagénomique virale ont été appliquées à différents types d'échantillons et de milieux (Figure In.7). Une grande partie des viromes s'est ainsi attachée à l'étude de communautés virales humaines, soit issues de différents tissus et organes, soit de manière moins invasive à partir de fluides issus du tractus digestif ou de l'appareil respiratoire. Des approches similaires ont également été appliquées aux animaux domestiques et sauvages, ainsi qu'aux végétaux. Les milieux aquatiques constituent l'autre grand type de milieu étudié, qu'il s'agisse de prélèvements océaniques, d'eau douce, ou d'environnement à forte salinité, acidité ou température. Enfin, plusieurs structures spécifiques ou milieux particuliers ont fait l'objet d'étude de métagénomique virale. Parmi ces structures, le corail est l'un des plus étudiés, notamment en raison de son importance écologique et des différentes atteintes sans cause déterminée dont les récifs coralliens sont victimes (Marhaver *et al.*, 2008; Vega Thurber *et al.*, 2008). Des analyses de viromes ont également été conduites sur des biofilms bactériens, comme des stromatolithes (Desnues *et al.*, 2008), des échantillons de nourriture en fermentation (Park *et al.*, 2011), ou encore des virus présents dans l'atmosphère terrestre (Whon *et al.*, 2012).

Caractéristiques générales des communautés virales

Les virus à ADN ont ainsi été étudiés dans des milieux très différents, et il est donc possible de dégager des tendances générales quant à ces communautés virales. La plupart des

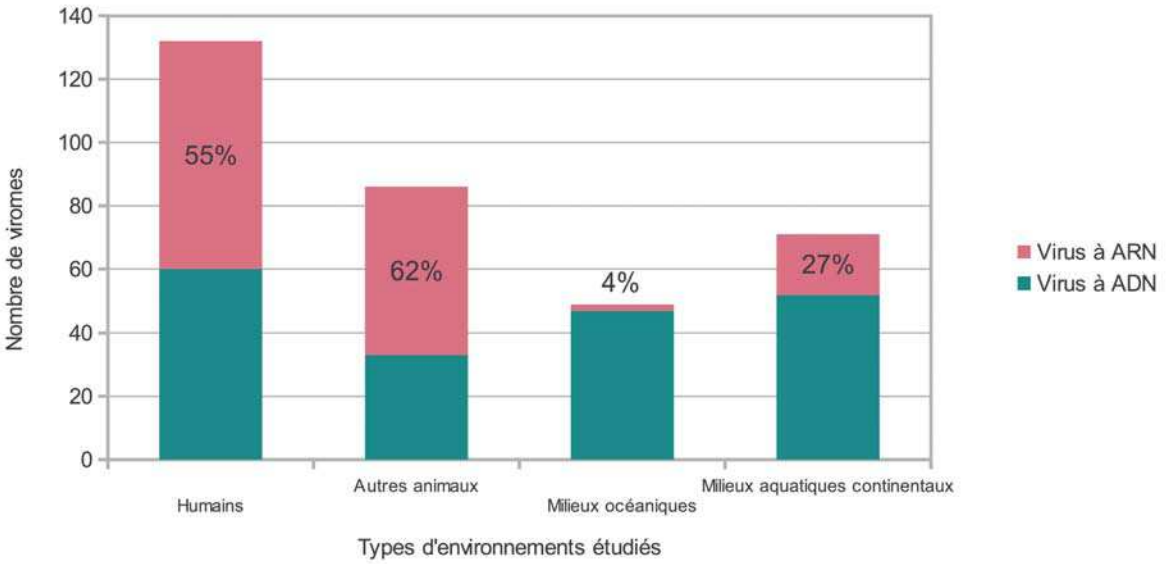


Figure In.7 : Nombre de viromes ciblant les communautés de virus à ADN ou ARN pour les quatre principaux types d'échantillons étudiés. *Le pourcentage de viromes ciblant les virus à ARN est indiqué pour chaque catégorie.*

analyses a tout d'abord révélé une très grande diversité génétique pour les virus à ADN, associée à une part très importante de séquences nouvelles, ne correspondant à rien de connu au sein des bases de données (Edwards & Rohwer, 2005). Ces résultats confirment les observations issues du séquençage de virus isolés, pour lesquels un nombre important de gènes non caractérisés avait été détecté. Quel que soit le milieu étudié, la fraction des séquences identifiée était le plus souvent majoritairement associée aux phages de type *Caudovirales*. Ce type de virus ayant été observé précédemment en microscopie électronique, les analyses de viromes ont ainsi confirmé leur caractère ubiquiste, et la grande diversité de leurs populations. Le deuxième groupe de virus fréquemment observé dans les viromes est celui des petits virus à ADN simple brin (Angly *et al.*, 2009; López-Bueno *et al.*, 2009; Rosario *et al.*, 2009a; Kim *et al.*, 2011). A l'inverse des *Caudovirales*, ce type de virus n'était pas détecté par les méthodes classiques d'observation (notamment en raison de leur taille), et les résultats de métagénomique ont véritablement révélé leur présence dans les différents milieux. De fait, ces virus restent encore largement à décrire du point de vue de leur diversité, leur distribution, leurs hôtes potentiel, et leurs liens évolutifs avec les autres groupes viraux et avec les génomes de leurs hôtes.

Cette multiplication des études portant sur les virus à ADN a permis de réaliser des méta-analyses, ou analyses comparatives intégrant un ensemble de données initialement analysées séparément. En 2009, Willner et collaborateurs ont ainsi montré par l'analyse de 45 métagénomes microbiens et 41 viromes qu'il semblait exister une signature propre à chaque type d'environnement au niveau des fréquences des motifs de deux nucléotides (ou dinucléotides) au sein des génomes (Willner *et al.*, 2009b). Cette similarité entre les communautés virales issues du même type d'environnement indiquerait ainsi soit une

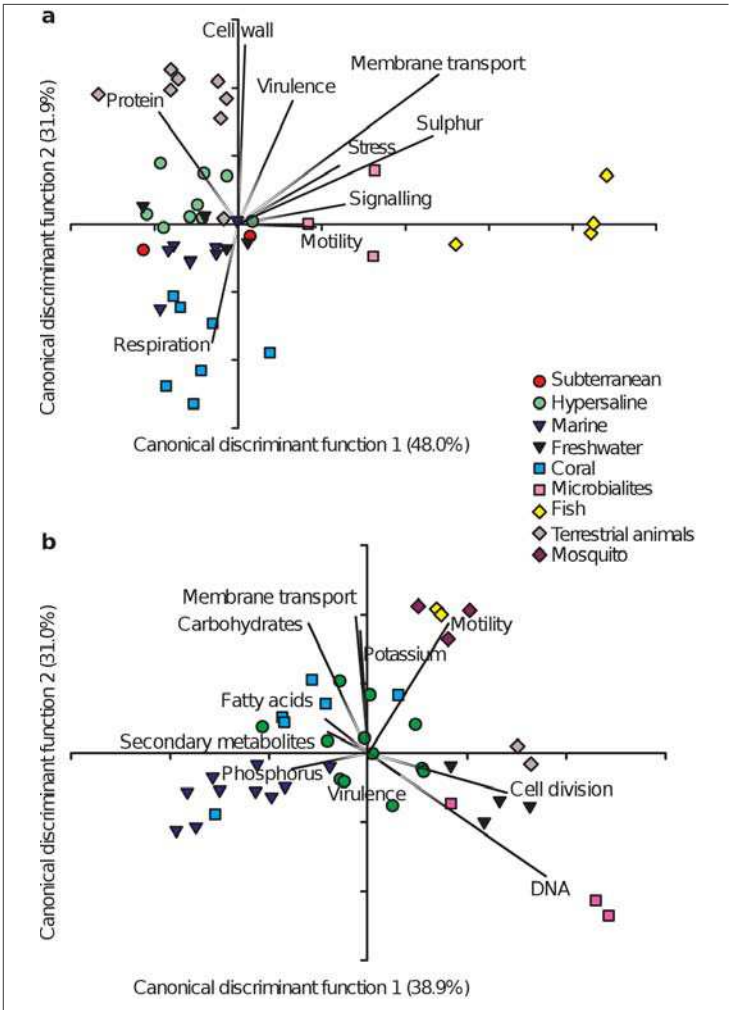


Figure In.8 : Analyse canonique des correspondances de la composition fonctionnelle de 45 microbiomes (A) et 42 viromes (B). *Figure issue de Dinsdale* et al., *2008a.*

évolution convergente sur le plan de la composition en nucléotides, soit l'existence d'échanges entre les différents points d'échantillonnage, pourtant physiquement séparés. À partir du même jeu de données, une conclusion similaire à été établie par Dinsdale et collaborateurs à partir de l'étude des potentiels fonctionnels des communautés microbiennes et virales (Dinsdale *et al.*, 2008). En effet, la comparaison de ces compositions fonctionnelles semble indiquer que le type d'écosystème est un élément structurant quant aux fonctions contenues dans les communautés (Figure In.8). De manière plus surprenante, un grand nombre de fonctions considérées comme typiquement cellulaires sont détectées au sein des communautés virales, laissant penser que les génomes viraux pourraient en réalité contenir beaucoup plus de gènes associés au métabolisme cellulaire que ce que les génomes complets laissaient supposer, et disposer ainsi d'un potentiel fonctionnel quasiment équivalent à celui des communautés microbiennes.

À l'inverse des virus à ADN, les viromes ciblant les virus à ARN sont majoritairement associés à des prélèvements humains et animaux, et il existe aujourd'hui peu d'études publiées concernant les populations environnementales de virus à ARN (Figure In.7). Ces quelques études indiquent que, si les virus à ARN de l'environnement sont majoritairement inconnus, la diversité au sein de ces communautés semble moins importante que celle des populations de virus à ADN. Cette richesse plus faible a d'ailleurs permis d'assembler de petits génomes complets, apportant plus d'information biologique que des séquences fragmentées pour la caractérisation de ces nouveaux types de virus (Culley *et al.*, 2006).

Ces études de métagénomique virale partagent donc un ensemble de méthodologies et de protocoles leur permettant d'étudier un vaste ensemble d'échantillons. Dans ce cadre, il est possible de distinguer trois grands types d'application de ces approches : les études purement cliniques, les études de communautés virales associées aux organismes supérieurs comme l'Homme, et enfin les études écologiques d'échantillons environnementaux.

Applications cliniques des viromes

Dans le cadre d'approches cliniques ou biomédicales dans lesquelles le virus est considéré comme un agent infectieux potentiel pouvant être à l'origine d'une pathologie, les viromes permettent une approche exploratoire, rendant possible la détection de nouvelles associations entre une série de symptômes et la présence de certains virus. Ces analyses ont surtout été réalisées sur les virus à ARN, puisqu'ils constituent l'essentiel des pathogènes viraux humains, et que ces génomes à ARN sont plus faciles à isoler que les génomes à ADN. Ces études de métagénomique virale associées à des organismes eucaryotes se sont toutefois

régulièrement heurtées à la grande richesse de ces communautés, qui a parfois rendu difficile tout établissement de lien entre la présence d'un virus et l'apparition de symptômes.

L'étude de Palacios et collègues, publié en 2008 constitue toutefois un exemple de découverte d'un nouveau virus *via* une approche métagénomique (Palacios *et al.*, 2008). Trois patients ayant reçu une greffe d'organe du même donneur ont tous contracté une affection fébrile aiguë, sans qu'une explication puisse être trouvée, puisque les tests de culture, en PCR, ou par puces à ADN n'ont pas révélé la présence d'un pathogène connu. Ainsi, les auteurs ont utilisé un séquençage massif et aléatoire des ARN isolés à partir d'organes du donneur : au sein des 100 000 séquences obtenues, 14 correspondaient à un nouveau type *d'Arenavirus.* La présence de ce virus a ensuite été confirmée dans l'ensemble des organes du donneur, tandis que les prélèvements effectués chez des individus sains ont tous été négatifs. Le lien avec les symptômes observés a également été confirmé par l'isolement et la caractérisation d'immunoglobuline ciblant ces *Arenavirus* dans l'un des patients, confirmant l'infection virale. Ces résultats confirment le potentiel exploratoire des viromes, mais mettent aussi en avant les différentes analyses nécessaires après l'étude du virome pour confirmer le lien éventuel entre les virus identifiés et les pathologies observées.

Ce type d'analyse a également été menée sur plusieurs échantillons en parallèle en s'appuyant sur des programmes de surveillance épidémiologique (Svraka *et al.*, 2010), ou sur des campagnes d'échantillonnage au niveau national (Victoria *et al.*, 2009) voire mondial (Finkbeiner *et al.*, 2008). Les approches métagénomiques ont par exemple été utilisées pour comprendre et surveiller certaines grandes pandémies récentes, comme les différentes épidémies de grippe (Nakamura *et al.*, 2009; Greninger *et al.*, 2010), pour lesquelles de nouveaux sous-types qui n'auraient pas été détectés par PCR ont pu être mis en évidence.

Caractérisation de la flore virale intestinale chez l'humain

Au-delà des applications cliniques, la métagénomique a rapidement été utilisée pour caractériser la flore virale associée au microbiote des organismes supérieurs. Dans ce cadre, la flore intestinale humaine a été le premier type d'échantillon étudié, avec la publication dès 2003 de plusieurs viromes issus de prélèvements de fèces (Breitbart *et al.*, 2003). Cette première étude de métagénomique virale sur la flore intestinale humaine a ainsi permis d'estimer que ces populations étaient constituées de 1 500 à 2 000 virotypes (génotypes viraux), les plus abondants représentant entre 2 et 3 % de la communauté.

Par la suite, plusieurs études comparatives ont permis de mieux comprendre les liens entre les individus et leur communauté microbienne et virale associée, ainsi que les modifications pouvant intervenir avec le temps et sous l'influence de l'alimentation (Reyes *et*

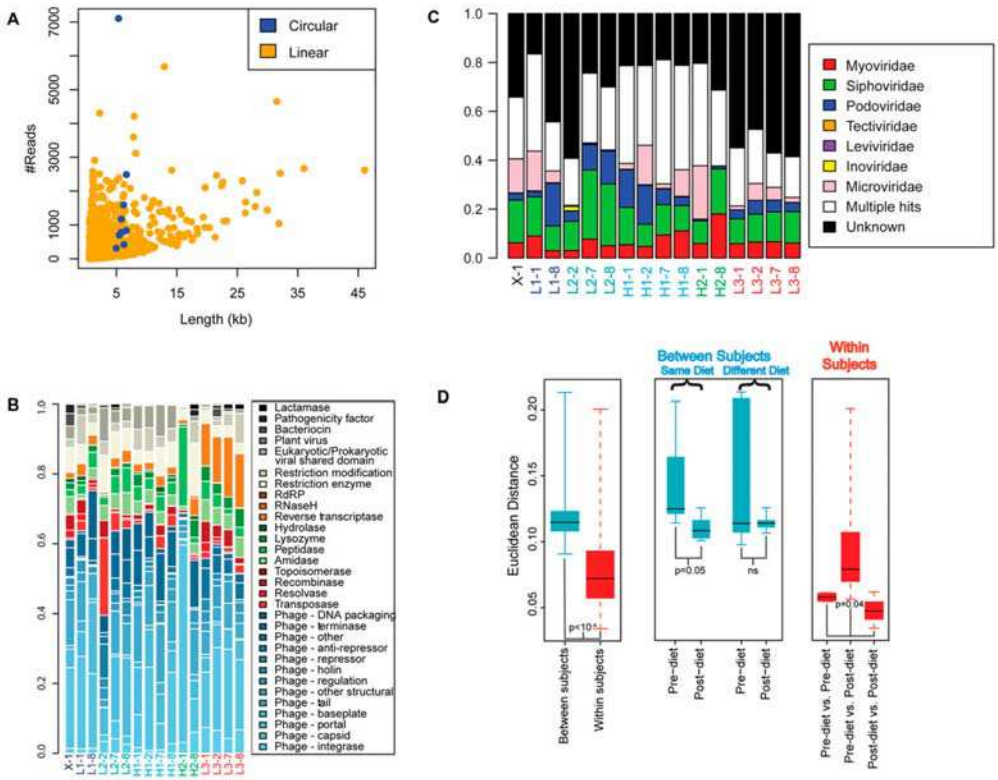


Figure In.9 : Analyse de viromes intestinaux humains avant et après un régime pour différents individus (adapté de Minot* et al., *2011). *A : Résultats de l'assemblage des viromes, associant la taille de chaque contig (en abscisse) avec son taux de couverture (en ordonnée). B : Affiliation fonctionnelle des gènes prédits pour chaque virome. C : Affiliation taxonomique pour chaque virome. D : Comparaisons entre viromes analysant les effets croisés du type de régime et de l'individu sur la flore virale intestinale.*

al., 2010; Minot *et al.*, 2011). Ces différentes études ont tout d'abord mis en avant la très forte spécificité des communautés virales intestinales entre individus, qui étaient de plus relativement stables dans le temps. La comparaison de microbiomes et de viromes issus de jumeaux monozygotes et de leurs ascendants directs (Reyes *et al.*, 2010) a ainsi montré que les communautés microbiennes semblaient similaires entre individus liés génétiquement, et qu'il existait au contraire une spécificité individuelle de la flore virale, y compris entre deux prélèvements séparés de plusieurs mois. De même, l'analyse de viromes issus d'individus soumis à différents régimes (Minot *et al.*, 2011) a montré que chaque individu portait une flore virale particulière, même si un changement drastique de régime alimentaire pouvait en modifier la structure et la composition (Figure In.9).

Plus récemment, Minot et collaborateurs ont présenté la première étude de métagénomique virale associée au séquençage de type HiSeq Illumina, permettant une profondeur de séquençage inédite (Minot *et al.*, 2012a, 2012b). Ce niveau de séquençage permet pour la première fois la reconstitution de génomes de phages complets (de type *Caudovirales*), avec en outre un nombre de séquences suffisant pour identifier les variations intra-population associées à chaque séquence consensus assemblée. L'analyse des loci

hypervariables au sein des contigs assemblés a ainsi révélé la présence en grande nombre de systèmes de génération de diversité par rétro-transcription puis recombinaison (similaire au système caractérisé dans le phage BPP1). Un large éventail de gènes était ciblé par ce système de génération de diversité, certains similaires à des gènes connus comme les immunoglobulines, et d'autres non caractérisés. Ainsi, les génome de bactériophages de la flore intestinale humaine comprennent de manière courante des mécanismes sophistiqués de mutation dirigée, et ces derniers ne constituent pas une exception dans un génome particulier. L'ensemble de ces études de viromes a ainsi permis de mieux caractériser les mécanismes d'évolution et la dynamique des populations au sein des virus à ADN de l'intestin humain.

Viromes de milieux océaniques

Les approches de métagénomique ont également été utilisées dans une optique écologique, visant à décrire les communautés virales de différents milieux, puis à comprendre la répartition spatiale des virus, identifier les facteurs régulant cette distribution, et déterminer leurs dynamiques temporelles.

Les premiers viromes aquatiques ont été générés à partir d'échantillons marins, au sein de la colonne d'eau ou à partir de sédiments (Breitbart *et al.*, 2002, 2004a). Ces jeux de données issus d'un séquençage de type Sanger ont permis d'estimer la richesse locale des communautés entre 300 et 10 000 virotypes différents par échantillon. La comparaison de ces premiers viromes aquatiques a notamment mis en évidence une part importante de séquences communes entre les deux types d'échantillon, confirmant ainsi le lien étroit entre les communautés virales de la colonne d'eau et celles de la surface des sédiments. Par la suite, l'utilisation des séquenceurs de nouvelle génération, et en premier lieu du pyroséquençage 454 de Roche (Angly *et al.*, 2006), révélera une diversité encore plus importante que ce que permettaient d'estimer les viromes précédents (jusqu'à 129 000 virotypes différents par échantillon).

Les études les plus récentes bénéficiant des développements méthodologiques (notamment en matière de purification des particules virales et au niveau des techniques de séquençage) ont pu effectuer des analyses comparatives d'un ensemble de viromes prélevés le long de transects au sein de l'océan Indien (Williamson *et al.*, 2012) ou Pacifique (Hurwitz & Sullivan, 2013). Ces analyses comparatives ont permis de détecter des gènes conservés entre les différents échantillons au sein des séquences connues comme inconnues, mais également permis d'acquérir des informations quand à la distribution des communautés virales dans les différentes masses d'eau. La richesse génétique des communautés virales semble ainsi

diminuer avec la profondeur de l'échantillon, avec la distance de l'échantillon par rapport à la côte, et enfin en période estivale (Hurwitz & Sullivan, 2013).

Des viromes ont également été réalisés à partir de prélèvements d'eau profonde, dans une baie (Steward & Preston, 2011) ou au sein de cheminées hydrothermales (Williamson *et al.*, 2008a). Dans les deux cas, au-delà du nombre important de séquences non identifiées, les deux études montrent qu'un grand nombre de virus de ces milieux profonds semblent capable d'établir un cycle lysogénique. Ainsi, les virus tempérés (lysogéniques) pourraient être majoritaires dans ces milieux, et impacter fortement les génomes de leur hôtes. La diversité de ces virus tempérés dans les milieux marins a également été étudiée au niveau des eaux de surface. Un virome a ainsi été généré à partir d'échantillon de bactéries marines après induction du cycle lytique. Ainsi, les génomes viraux étudiés sont majoritairement issus de virus tempérés (McDaniel *et al.*, 2008). À nouveau, le taux de séquences identifiées au sein de ce virome est bas (6 %), et la diversité génétique observée est très élevée. Ces résultats semblent bien indiquer qu'un nombre très important de virus marins sont capables d'établir un cycle lysogénique.

Viromes de milieux aquatiques continentaux

Les approches de métagénomique virale ont également été appliquées à différents types de milieux aquatiques continentaux, le plus souvent dans le cadre d'études ponctuelles de milieux très spécifiques.

Les milieux aquatiques continentaux impactés par l'Homme ont fait l'objet d'une attention toute particulière, notamment concernant l'origine des virus retrouvés dans ces milieux. Par exemple, Rosario et collaborateurs rapportent en 2009 l'analyse de viromes issus de différents points au sein d'une station d'épuration. Ces différents viromes semblent montrer que les eaux usées peuvent servir de réservoir de virus infectant les micro-organismes, et potentiellement les plantes et animaux, mais aucun virus humain ne sera détecté (Rosario *et al.*, 2009b). Des résultats similaires seront obtenus en 2012 par l'étude de viromes issus d'eaux usées non traitées, avec cette fois la présence de familles virales infectant potentiellement les humains (Ng *et al.*, 2012).

L'analyse de lacs hypersalins a également été menée, et a mis en avant la spécificité très marquée de ces communautés virales halophiles (Santos *et al.*, 2010; Emerson *et al.*, 2012; Garcia-Heredia *et al.*, 2012). Ces différentes études semblent notamment faire état d'une diversité génétique limitée pour les communautés de virus à ADN de ces milieux hypersalins. En 2012, Emerson et collaborateurs rapportent ainsi l'assemblage de génomes complets allant jusqu'à 60 kb à partir de séquences de viromes, conséquences de la richesse

relativement modérée de ces milieux et de l'évolution des techniques de séquençage, ces viromes bénéficiant des dernières versions du séquenceur HiSeq d'Illumina.

D'autres milieux “extrêmes”, notamment fortement acide et à haute température, ont aussi fait l'objet d'analyses de métagénomique virale. En 2008, Schoenfeld et collaborateurs décrivent ainsi deux viromes issus de sources hyperthermales du parc de Yellowstone (Schoenfeld *et al.*, 2008). Parmi les conclusions de l'étude, les auteurs notent en particulier un nombre important de nouveaux gènes, et considèrent que les organismes des milieux aquatiques continentaux extrêmes, et particulièrement les virus, sont vraisemblablement porteurs de gènes et d'enzymes entièrement nouveaux, aux capacités potentiellement intéressantes d'un point de vue biotechnologique (Schoenfeld *et al.*, 2010). En 2012, l'analyse d'un virome issu d'un lac hyperacide et à forte température a révélé l'existence d'un virus issu d'une recombinaison entre un virus à ARN et un virus à ADN, confirmant que ces environnements extrêmes abritent certainement des communautés virales très spécifiques, potentiellement porteuses innovations génétiques (Diemer & Stedman, 2012).

Les milieux d'eau douce étudiés jusqu'à maintenant sont eux aussi très singuliers. La première analyse de métagénomique virale appliquées aux communautés de virus à ADN d'un lac d'eau douce a présenté en 2009 la comparaison de deux échantillons du lac Limnopolar, situé en Antarctique, et dont la surface est recouverte de glace neuf mois par an (López-Bueno *et al.*, 2009). Les communautés virales sont à nouveau majoritairement non caractérisées, mais des différences importantes entre les deux échantillons ont été identifiées, notamment au

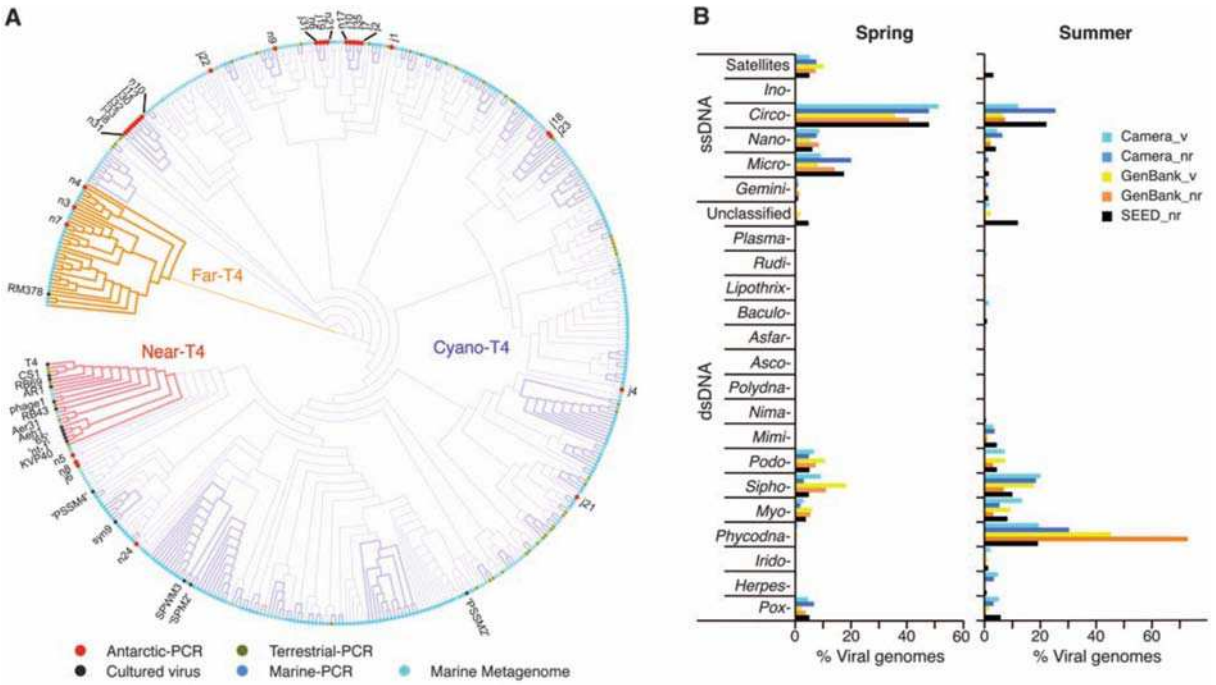


Figure In.10: Analyse de viromes lacustres issus du lac Limnopolar (Antarctique, figure tirée de Lopez Bueno et al., 2009). *A : Arbre phylogénétique basé sur la protéine majeure de capside pour les phages de type T4. B : Composition taxonomique des fractions identifiées des viromes du lac Limnopolar au printemps (lac gelé) et en été (lac dégelé).*

niveau des virus du phytoplancton aquatique (Figure In.10). En effet, les virus infectant le phytoplancton ne sont pas détectés au sein du premier virome lorsque le lac est encore gelé, mais constituent l'essentiel de la communauté virale identifiée au sein du deuxième virome (absence de couverture de glace). Ce résultat pointe ainsi l'origine probablement allochtone de ces virus impliqués dans la régulation des efflorescences phytoplanctoniques au sein de ces lacs. Plus récemment, Fancello et collaborateurs ont publié l'analyse de viromes issus de quatre points d'eau au sein du Sahara (Fancello *et al.*, 2012b). Cette étude a notamment mis en évidence une proportion importante de virus lysogéniques pour un point d'eau aux conditions particulièrement extrêmes (niveau d'eau très bas). Cet exemple illustrerait ainsi l'adaptation de la communauté virale aux conditions environnementales dans lesquelles évoluent leurs hôtes.

Une seule étude, ciblant la communauté de virus à ARN, s'est intéressée à un lac de région tempéré (Djikeng *et al.*, 2009). De manière surprenante, une part importante des séquences identifiées sont associées aux virus à ADN, ce qui laisse penser que la séparation des deux types de génomes a été difficile pour ces échantillons. La plupart des virus retrouvés semblent associés à des hôtes qu'on ne trouve pas dans les lacs (principalement des vertébrés, dont des oiseaux et porcins). Les lacs pourraient ainsi, selon les auteurs, constituer un réservoir de virus pour lesquels ces milieux ne seraient qu'un lieu de passage.

Enfin, Rodriguez-Brito et collaborateurs ont étudié les communautés virales d'un ensemble de lacs, depuis l'eau douce jusqu'à des bassins hypersalins (Rodriguez-Brito *et al.*, 2010). Si les virus retrouvés semblent différents entre les types d'échantillons, la dynamique saisonnière leur est toutefois apparue comme très semblable, avec une stabilité des communautés sur un temps long (plusieurs mois), et des variations plus importantes à court terme (quelques jours) et à des niveaux taxonomiques très fins (génotype viral ou souche bactérienne). Toutefois, les conclusions obtenues à partir de ces données sont fortement limitées par le faible nombre de séquences par échantillon et leur taille (100 pb).

En conclusion, le métagénomique appliquée à l'étude des virus constitue une approche totalement nouvelle de ces organismes, et a déjà permis de mieux identifier les différents acteurs des communautés virales, qu'elles soient associées à des eucaryotes ou au sein d'écosystèmes naturels. En particulier, un grand nombre de nouvelles souches virales ont pu être caractérisées grâce à ces viromes, et parfois associées soit à certaines pathologies, soit à des paramètres environnementaux. La discipline est cependant encore très jeune, et les méthodes utilisés pour la préparation et l'analyse de ces viromes ne sont pas encore standardisées. En particulier, les outils d'analyse bioinformatique dédiés à la métagénomique virale sont en grande partie à développer ou en cours de développement. De même, si le nombre d'études portant sur le milieu océanique semble

comparable au nombre d'études portant sur les milieux aquatiques continentaux, ces derniers couvrent un panel d'environnements et de biomes beaucoup plus important. Il existe ainsi un manque important d'information au sujet des communautés de virus à ARN des milieux aquatiques et des virus à ADN des milieux lacustres tempérés. De même, les analyses de viromes aquatiques sont généralement ponctuelles et restreintes à un milieu spécifique.

Objectifs de travail et présentation des travaux réalisés

Ce travail de doctorat s'est ainsi constitué autour de l'analyse bioinformatique de métagénomes viraux, du développement d'outil aux méta-analyses, afin de mieux caractériser **les virus de l'environnement, depuis l'échelle des communautés virales jusqu'à l'analyse de génotypes spécifiques**.

Un premier chapitre traitera des différents outils d'analyse développés, regroupés au sein du premier serveur web dédié à l'analyse des viromes : **Metavir**. L'objectif de ce serveur est de mettre à la disposition des biologistes les différents logiciels existants ainsi que de nouveaux outils développés, adaptés aux spécificités des métagénomes viraux. Une seconde série d'outils (**Metavir 2**) a par la suite été conçue pour répondre aux évolutions rapides des techniques de séquençage à haut-débit, et en particulier au changement de type de données étudiées (des séquences brutes aux séquences assemblées).

Les métagénomes viraux seront ensuite étudiés en tant que fenêtre ouverte sur le pangénome encapsidé. Différentes analyses de viromes seront ainsi présentées, centrées tout d'abord sur le potentiel fonctionnel des génomes viraux, puis orientées autour de la diversité virale dans les milieux aquatiques, en termes de gènes et de génotypes.

Le deuxième chapitre sera ainsi consacré au **potentiel fonctionnel des communautés virales** environnementales tel qu'estimé à partir des viromes. Des différences importantes ont en effet pu être notées entre l'estimation de ce potentiel fonctionnel par les approches métagénomiques, et cette même estimation à partir des génomes viraux complets. Nous avons ainsi cherché à comprendre les raisons de ces différences, et s'il était possible de concilier ces deux visions du contenu global des génomes viraux.

Un troisième chapitre tentera d'analyser la distribution des gènes et génotypes viraux au sein des milieux aquatiques autour du globe, et d'en comprendre les paramètres et facteurs d'influence. Une première analyse s'est attachée à décrire les **communautés virales de milieux lacustres**, pour lesquelles très peu de données étaient disponible. Une seconde étude s'est focalisée sur les facteurs influençant la distribution des différentes espèces virales au sein

des différents milieux aquatiques. En considérant que la salinité a souvent été notée comme structurant les communautés de micro-organismes, une série de viromes a été réalisée le long d'un **gradient de salinité,** afin d'analyser les liens éventuels entre paramètres physico-chimiques, distance géographique entre prélèvements, et composition de la communauté virale.

Enfin, un dernier chapitre est consacré au petits virus à ADN simple brin, éléments génétiques simples, souvent utilisés comme organismes modèles en laboratoire, mais dont la diversité et les mécanismes d'évolution sont encore méconnus. Par une approche d'assemblage d'un ensemble de viromes, une collection de nouveaux génomes complets a pu être obtenue pour deux familles de petits virus à ADN simple brin. Cette procédure a tout d'abord été appliquée aux ***Microviridae***, famille de phages à ADN simple brin régulièrement détectée dans différentes analyses de viromes, des milieux aquatiques au tractus digestif humain. La même méthodologie a ensuite été utilisée pour étudier la diversité des **virus chimères**, groupe de virus récemment décrits issus de la recombinaison entre un virus ARN simple brin et un virus à ADN simple brin.

À travers ces différents travaux, nous avons ainsi cherché à mieux caractériser la diversité génétique et génomique des communautés virales environnementales, et ainsi révéler différents aspects encore inconnus de la virosphère tels que les facteurs structurant les communautés virales, leur potentiel métabolique, et les mécanismes évolutifs à l'œuvre. Ces analyses ont nécessité le développement d'outils bioinformatiques, que nous avons choisi de rendre disponible à travers Metavir, le premier serveur web dédié à l'étude de viromes.

Chapitre I- Développement d'outils bioinformatiques dédiés à l'analyse de viromes : le serveur web Metavir

Les approches métagénomiques ont été développées durant les deux dernières décennies, et se sont rapidement imposées dans le cadre de l'écologie microbienne, notamment *via* la réalisation de projets de grande ampleur incluant plusieurs centaines d'échantillons (Yooseph *et al.*, 2007; Karsenti *et al.*, 2011; Schloissnig *et al.*, 2013). La mise au point d'une série de protocoles expérimentaux visant à extraire les capsides virales et le matériel génétique encapsidé a par la suite permis d'appliquer ce type d'approches aux communautés virales (Casas & Rohwer, 2007; Vega Thurber *et al.*, 2009a). Ainsi, et même s'il reste certains développements méthodologiques à réaliser en terme de purification des capsides virales à partir d'un échantillon, l'analyse des données s'est rapidement révélée être l'étape charnière des études de métagénomique virale plus que l'obtention de ces données.

En effet, même si les serveurs généralistes dédiés au traitement de données métagénomiques comme Mg-Rast (Meyer *et al.*, 2008) et Camera (Sun *et al.*, 2011) permettent de réaliser l'analyse complète de métagénomes microbiens, ils ne sont toutefois pas adaptés aux données virales. Tout d'abord, les deux sites s'appuient en partie sur les séquences similaires aux gènes codant pour les ARN ribosomaux afin de déterminer la composition taxonomique de l'échantillon, or ces gènes sont absents des génomes viraux. Une telle analyse globale à partir d'un seul gène marqueur n'est d'ailleurs pas transposable aux viromes puisqu'aucun gène n'est conservé chez l'ensemble des virus. De plus, les différentes affiliations des séquences métagénomiques, qu'elles soient fonctionnelles ou taxonomiques, sont utilisées pour comparer ces jeux de données. Or pour une écrasante majorité de métagénomes viraux, le taux de séquences affiliées est très bas (généralement inférieur à 30%). Ainsi, dans le cas des viromes, ces comparaisons se baseraient sur une minorité de séquences et seraient fortement biaisées.

Ces spécificités des génomes viraux, en particulier l'absence de gène marqueur universel et le faible taux de séquences affiliées, ont entraîné le développement d'outils bioinformatiques spécifiquement adaptés aux données de métagénomique virale. Dès 2005, Angly et collaborateurs proposent le logiciel PHACCS, afin d'estimer la structure d'une population virale à partir de l'assemblage d'un métagénome (Angly *et al.*, 2005). Plusieurs outils visant à estimer la composition taxonomique à partir d'un virome ont également été publiés, comme GAAS (Angly *et al.*, 2009), et ProVide (Ghosh *et al.*, 2011). Toutefois, aucun de ces logiciels ne propose une analyse générale et complète d'un métagénome viral, et ils sont pour la plupart destinés à être installés sur un ordinateur de manière locale.

Dans ce contexte, nous avons décidé de mettre en place un serveur web dédié à l'analyse des viromes : Metavir, proposant un ensemble d'outils pour partie déjà publiés et

pour partie développés spécifiquement pour le serveur, et visant également à centraliser un ensemble de viromes publiés.

Metavir : un serveur web pour l'analyse des viromes

La première version du serveur Metavir, mise en ligne en septembre 2011, comportait quatre outils : deux dédiés aux séquences similaires aux séquences de virus connus, et deux outils ayant pour but une caractérisation et comparaison des jeux de données dans leur ensemble. Dès la mise en ligne, 44 viromes provenant de différents types de biomes étaient publiquement disponibles, constituant ainsi une base de viromes de référence.

Support matériel

Le serveur Metavir est basé sur une architecture linux (distribution CentOS), et adossé à un cluster de calcul de 48 nœuds (agrandi par la suite). L'ensemble de cette infrastructure est géré et administré par le CRRI (Centre régional de ressources informatiques ; http://crri.clermont-universite.fr/). Perl et R sont les deux principaux langages utilisés dans la conception des outils de Metavir, tandis que la partie web a été réalisée en associant PHP, CSS et JavaScript. L'ensemble de ces scripts et programmes s'articulent autour d'une base de donnée MySQL.

Affiliation des séquences de viromes

Au sein de cette première version de Metavir, deux outils sont dédiés à l'affiliation des séquences. Dans un premier temps, une composition taxonomique est déduite à partir des meilleurs hits du BLAST comparant les séquences du viromes avec la base de données RefseqVirus, composée de l'ensemble des génomes viraux complets. Les affiliations par meilleur résultat de BLAST sont ensuite normalisées par la taille moyenne du génome grâce à l'outil GAAS (Angly *et al.*, 2009), afin d'obtenir une composition reflétant au mieux le nombre de virions de chaque virus présent dans l'échantillon.

S'il est possible d'obtenir une image de la composition taxonomique d'un échantillon (du moins de sa partie identifiée) grâce à une comparaison par BLAST, une affiliation plus fine peut être réalisée *via* l'utilisation d'outils phylogénétiques. Afin de pouvoir utiliser différents gènes marqueurs, une procédure a été développée pour générer automatiquement des arbres phylogénétiques à partir de séquences de viromes et d'un alignement de référence.

Comme les séquences métagénomiques sont souvent trop courtes pour contenir le gène complet, les alignements sont généralement partiels, et il est impossible d'obtenir un alignement multiple de bonne qualité comportant toutes les séquences métagénomiques. Ainsi, la zone d'alignement de chaque séquence de virome est automatiquement détectée, et utilisée pour regrouper les différentes séquences métagénomiques de manière à générer (quand cela est possible) différents arbres phylogénétiques comportant le maximum de séquences environnementales.

Huit gènes marqueurs (G20, GP23, TerL, T7gp17, PolB, MCP, VP1 et RC-Rep) étaient proposés lors de la première mise en ligne du serveur Metavir, couvrant certains des groupes majeurs comme les *Caudovirales* et les *Phycodnavirus*, ainsi que plusieurs familles de petits virus à ADN simple brin.

Structure et comparaison des viromes complets

Deux autres analyses étaient disponibles lors de la première mise en ligne de Metavir : l'une pour estimer la richesse génétique d'un virome, et l'autre pour réaliser une comparaison globale des jeux de données. La richesse génétique au sein d'un virome est évaluée *via* une clusterisation des séquences, proposée à différents seuils de similarité. Le nombre de clusters formé augmentera avec la richesse génétique d'un échantillon. Toutefois il n'est pas possible à partir de ces clusterisations de déterminer s'il s'agit de génomes complexes de grande taille, ou d'une multitude de petits génomes, les deux situations menant à la même observation. Ces clusterisations sont présentées sous la forme de courbes de raréfaction, ce qui permet d'observer la couverture par le virome de l'ensemble des gènes viraux initialement présents au sein de l'échantillon. En effet, si l'ensemble des gènes sont séquencés, la courbe atteindra un plateau car l'ajout de nouvelles séquences n'entraînera pas la création de nouveaux clusters. Afin de pouvoir comparer différents jeux de données comprenant des tailles de séquences différentes, ces analyses sont également proposées à partir de sous-échantillons normalisés (50 000 séquences de 100 pb).

Le dernier outil permet de comparer les viromes en utilisant à nouveau l'ensemble des séquences disponibles, et non uniquement les séquences affiliées. En effet, la fraction de séquences affiliées étant généralement faible, la comparaison classique utilisée pour les métagénomes bactériens ou eucaryotes se basant sur des résultats d'affiliations taxonomiques ou fonctionnelles ne peut être considérée comme exhaustive et robuste. Nous avons choisi d'implémenter un outil de comparaison basé sur une comparaison des séquences des viromes deux à deux, méthodologie inspirée d'une comparaison de métagénomes microbiens précédemment publiée (Martín-Cuadrado *et al.*, 2007). Ces scores de similarité entre viromes

sont répertoriés dans une matrice, sur laquelle sera basée un clustering (méthode du lien moyen). Ainsi, il est possible de positionner au sein d'un dendrogramme les différents jeux de données choisis qui seront d'autant plus proches que leurs séquences sont similaires, que ces séquences soient caractérisées ou non.

Article I

METAVIR: a web server dedicated to virome analysis

Simon Roux[1,2*], Mickael Faubladier[1,2], Antoine Mahul[3], Aurélien Bernard[1,2], Didier Debroas[1,2] and François Enault[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand
[2]CNRS, UMR 6023, LMGE, F-63177 Aubière
[3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France

Category : Data and Text Mining
Associate Editor: Alfonso Valencia

Publié le 11 septembre 2011 dans **Bioinformatics** (27 , 21 : 3074-3075)

BIOINFORMATICS APPLICATIONS NOTE Vol. 27 no. 21 2011, pages 3074–3075
doi:10.1093/bioinformatics/btr519

Data and text mining

Advanced Access publication September 11, 2011

Metavir: a web server dedicated to virome analysis

Simon Roux[1,2,*], Michaël Faubladier[1,2], Antoine Mahul[3], Nils Paulhe[1,2], Aurélien Bernard[1,2], Didier Debroas[1,2] and François Enault[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand, [2]CNRS, UMR 6023, LMGE, F-63177 Aubiere and [3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France

Associate Editor: Alfonso Valencia

ABSTRACT

Summary: Metavir is a web server dedicated to the analysis of viral metagenomes (viromes). In addition to classical approaches for analyzing metagenomes (general sequence characteristics, taxonomic composition), new tools developed specifically for viral sequence analysis make it possible to: (i) explore viral diversity through automatically constructed phylogenies for selected marker genes, (ii) estimate gene richness through rarefaction curves and (iii) perform cross-comparison against other viromes using sequence similarities. Metavir is thus unique as a platform that allows a comprehensive virome analysis.

Availability: Metavir is freely available online at:
http://metavir-meb.univ-bpclermont.fr

Contact: simon.roux@univ-bpclermont.fr

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on May 6, 2011; revised on July 29, 2011; accepted on August 29, 2011

1 INTRODUCTION

Describing environmental viral communities is a major challenge for environmental microbiology. Viruses are known to intervene in a broad spectrum of processes spanning population regulation, horizontal gene transfer and major biogeochemical cycles (Suttle, 2007). However, the study of environmental viral communities is made difficult by the fact that only a tiny fraction of their hosts has been cultivated. Moreover, as no single gene is common to all viral genomes, environmental viral communities cannot be studied via approaches based on ribosomal RNA sequencing (Edwards and Rohwer, 2005). One way around these limitations is to directly sequence the viral communities. Such metagenomic approaches can provide insights into the viral diversity of environments of interests and are progressively gaining wider uses (Allen and Wilson, 2008). Existing bioinformatics tools implemented on web servers dedicated to metagenome analyses are not specific to particular biological entities (Meyer *et al.*, 2008; Seshadri *et al.,* 2007). Yet, viral metagenomes (viromes) are by nature significantly different from bacterial metagenomes. Indeed, very little is known about viral communities and the strong sequence divergence and broad gene richness found in viromes indicate a tremendous genetic diversity (Kristensen *et al.,* 2009). Yet this diversity remains very poorly characterized as the databases lack annotated viral sequences. Most virome sequences (i.e. reads) therefore differ from previously described sequences (Edwards and Rohwer, 2005). However, direct comparisons of virome sequences from disparate locations often exhibit significant overlap, indicating that, although our knowledge on viral genes remains sparse, the same genes are seen everywhere (Polson *et al*. 2011). The unknown fraction of the reads (between 65% and 95%) usually left aside in classical metagenomic analyses represents extremely valuable data for virome analysis. Furthermore, different marker genes describing the different viral families are needed to further describe viral diversity and none of these markers are available on existing 'generalist' web servers.

To our knowledge, PHACCS (Phage Communities from Contig Spectrum; Angly *et al*., 2005) is the only tool designed for viral community sequencing, but it focuses on assessing the structure of uncultured viral communities in terms of ecological parameters.

Here we present Metavir, an interactive web server that performs a comprehensive analysis of viromes. Classical analyses are available (sequence characteristics, taxonomic composition). Marker genes selected for each major viral families can be used in a custom-designed procedure to perform phylogenetic analysis on virome reads. The automatically generated phylogenetic trees allow biologists to explore the viral diversity in a deep and precise manner. Finally, specific tools have been developed to efficiently deal with the vast unknown fraction. The gene richness of a virome can be assessed and compared to other viromes, and viromes can also be compared in terms of sequence similarity.

2 METHODS

Metavir provides users with a suite of tools inside a private environment for analyzing viromes. After a registration step, the user can submit viromes as fasta files. Metavir separates virome analysis into four major tools, as illustrated in the Supplementary Material using the Sargasso Sea virome (Angly *et al*., 2009).

2.1 Virome composition

Virome composition is assessed using the GAAS tool (Angly *et al*., 2009). Virome reads are compared to complete viral genomes from the Refseq database, and taxonomic affiliation results are normalized by genome length in order to estimate the number of viral particles for each viral species in the initial sample (Supplementary Fig. S1).

*To whom correspondence should be addressed.

2.2 Automatic phylogenies for marker genes

A procedure has been developed to insert metagenomic reads in phylogenetic trees containing reference sequences for chosen marker genes. This is the first automatic phylogeny generation procedure available for virome sequences. Phylogenies are of great utility to virome research due to the tremendous diversity observed in viral sequences and the lack of representative sequences in the databases. If enough reads are homologous to the marker gene, phylogenies can be generated from 100 bp reads but the procedure provides even better results with 400 bp reads commonly generated today by NGS tools such as 454 TITANIUM.

For each marker available, reference sequences have been retrieved from the PFAM database and aligned using MUSCLE (Edgar, 2004). A BLASTx is computed to detect potential homologous sequences in the virome, and all metagenomic reads having a BLAST hit against one of the reference sequences with an E-value $< 10^{-3}$ are gathered. These sequences are then compared to NR (BLASTx), and excluded from the analysis if their best BLAST hit does not correspond to the studied marker. The remaining reads are assembled using Cap3 (Huang and Madan, 1999) (98% identity on 35 bp) to be able to work with longer sequences. These parameters should only group sequences from the same virotype (Angly *et al*., 2005). These sequences are translated into protein sequences and then aligned against the reference alignment via a HMM profile using HMMER (Eddy, 1998). In order to generate trees containing several metagenomic sequences, alignment bounds for each metagenomic sequence are collected and used to define multiple subalignments. Alignments are cleaned using Gblocks (Talavera and Castresana, 2007) and used to generate phylogenetic trees with 100 bootstraps using PhyML (Guindon *et al*., 2009). Finally, the tree is rooted and monophyletic groups are highlighted via Scriptree (Chevenet *et al*., 2010) (Supplementary Fig. S2). This analysis is already available for the main viral families through different marker genes, such as VP1 for *Microviridae*, or TerL for *Caudovirales.* Users can request specific marker genes using a form on the website.

2.3 Virome comparison

In order to compare viromes in their entirety rather than only their small known fraction, a qualitative comparison of viromes based on sequence similarity (tBLASTx comparison) is computed as described in a previous work on bacterial metagenomes (Martín-Cuadrado *et al*., 2007). Virome samples of 50 000 sequences of 100 bp are used in order to have comparable results. Each sample is compared to every other sample using tBLASTx. A similarity score between virome A and virome B is then computed as the sum of best BLAST hit scores of virome A reads against virome B reads. Finally, the resulting score matrix (i.e. similarity scores for all virome pairs) is used to cluster viromes using R software and the pvclust package, working with default parameters and 100 bootstraps (Suzuki and Shimodaira, 2006). Users can choose to compare any virome subsets (Supplementary Fig. S3).

2.4 Rarefaction curves

Here again utilizing the whole virome rather than only the reads identified by BLAST, the rarefaction curves are computed to assess the gene richness of the viromes (Raes and Bork, 2008). Viromes can then be compared using this view of the genetic diversity within a viral community. Clusters are computed with Uclust (Edgar, 2010) and three thresholds are proposed: 75, 90 and 98%. A Perl script counts the number of different clusters generated for a given number of input sequences, in order to plot rarefaction curves. These rarefaction curves are computed on whole viromes (Supplementary Fig. S4), which makes it possible to determine whether the entire gene pool is sampled in the virome (in which case the rarefaction curve would level off), and on virome samples (50 000 sequences of 100 bp), in order to compare gene richness between different ecosystems (Supplementary Fig. S5). Users can select any virome subset to be plotted into rarefaction curves. Curves are dynamically generated with JS Charts.

3 WEB INTERFACE AND IMPLEMENTATION

A set of previously published viromes has already been included in Metavir as public projects available for any user (registered or not). Private projects are restricted to the user who uploaded the project, until such time as the user decides to make it public. Metavir computations are distributed on a cluster allowing multiple parallel runs (40 CPU). A full analysis of a large-sized virome (600 000 reads of 400 pb) would take a few days in total.

Funding: This work was supported by the French 'Defence Procurement Agency' (DGA).

Conflict of Interest: none declared.

REFERENCES

Allen,M.J. and Wilson,W.H. (2008) Aquatic virus diversity accessed through omic techniques: a route map to function. *Curr. Opin. Microbiol.,* **11**, 226–232.

Angly,F. *et al*. (2005) PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*, **6**, 41.

Angly,F. *et al*. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput. Biol*, **5**, e1000593.

Chevenet,F. *et al*. (2010) ScripTree: scripting phylogenetic graphics. *Bioinformatics*, **26**, 1125–1126.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.

Edgar,R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Edwards,R.A. and Rohwer,F. (2005) Viral metagenomics. *Nat. Rev. Micro.*, **3**, 504–510.

Guindon,S. *et al*. (2009) Estimating maximum likelihood phylogenies with PhyML. *Methods Mol. Biol.*, **537**, 113–137.

Huang,X. and Madan,A. (1999) CAP3: a DNA Sequence Assembly Program. *Genome Res.*, **9**, 868–877.

Kristensen,D.M. *et al.* (2010) New dimensions of the virus world discovered through metagenomics. *Trends Microbiol.,* **18**, 11–19.

Martín-Cuadrado,A.-B. *et al*. (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE*, **2**, e914.

Meyer,F. *et al*. (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.

Polson,S.W. *et al*. (2011) Unraveling the viral tapestry (from inside the capsid out). *ISME J.*, **5**, 165–168.

Raes,J. and Bork,P. (2008) Molecular eco-systems biology: towards an understanding of community function. *Nat. Rev. Microbiol.,* **6**, 693–639.

Suttle,C.A. (2007) Marine viruses—major players in the global ecosystem. *Nat. Rev. Micro.*, **5**, 801–812.

Suzuki,R. and Shimodaira,H. (2006) Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics*, **22**, 1540–1542.

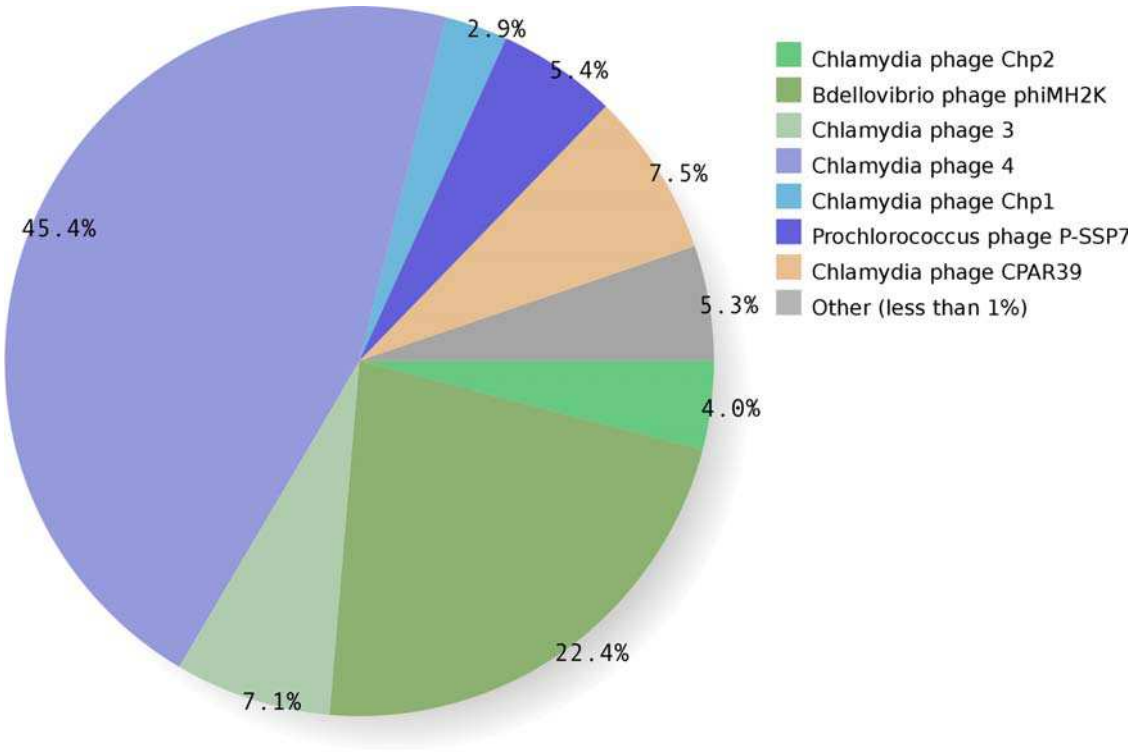
Seshadri,R. *et al.* (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.,* 2007, **5**, e75.

Talavera,G. and Castresana,J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.*, **56**, 564–577.

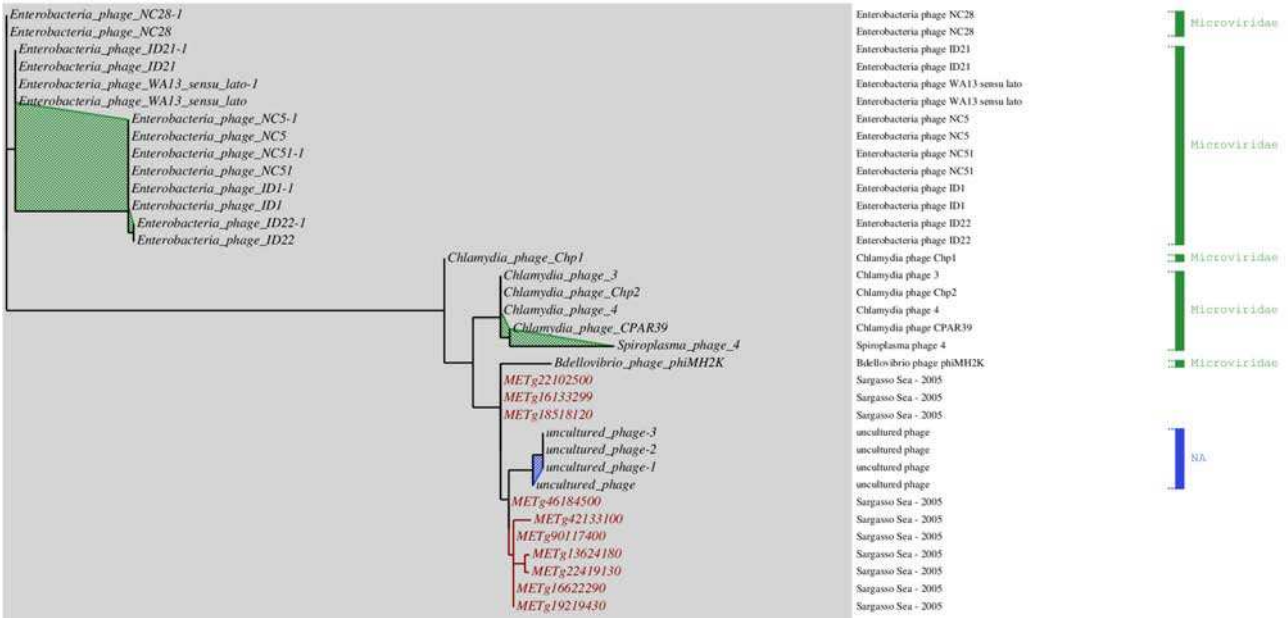
Supplementary Data

Usage example of Metavir

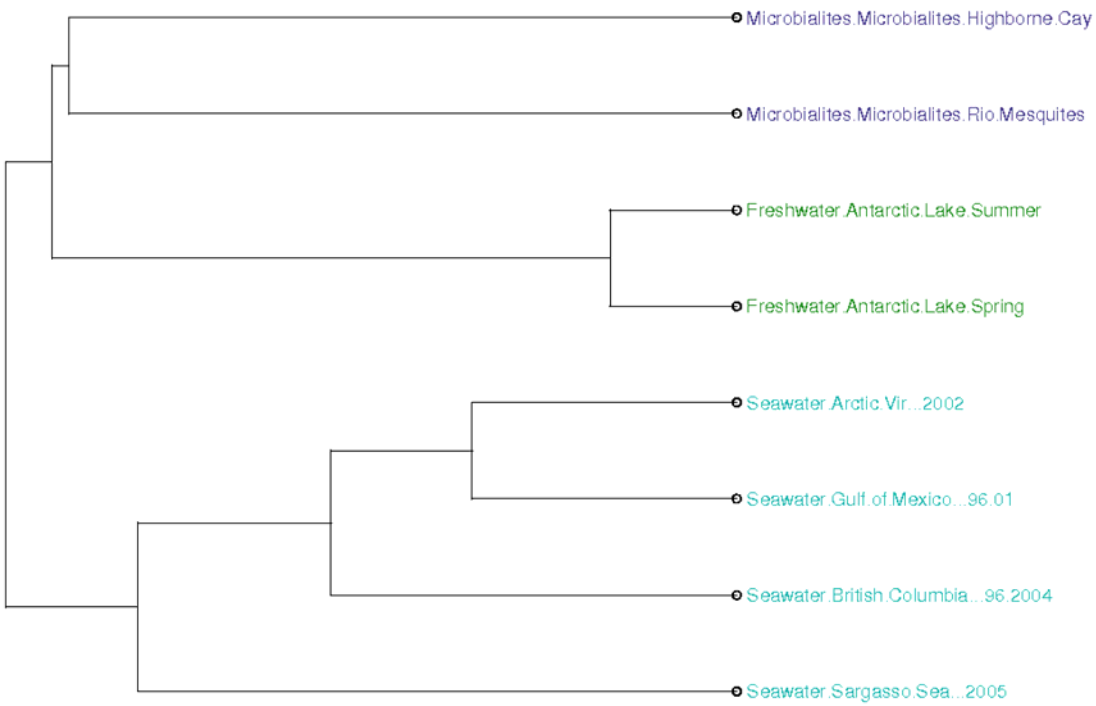
This supplement provides further details on the major tools available in the METAVIR web server, through the analysis of a real data set, the Sargasso Sea virome (Angly *et al.,* PLoS Comput Biol. 2009 December; 5(12): e1000593.).



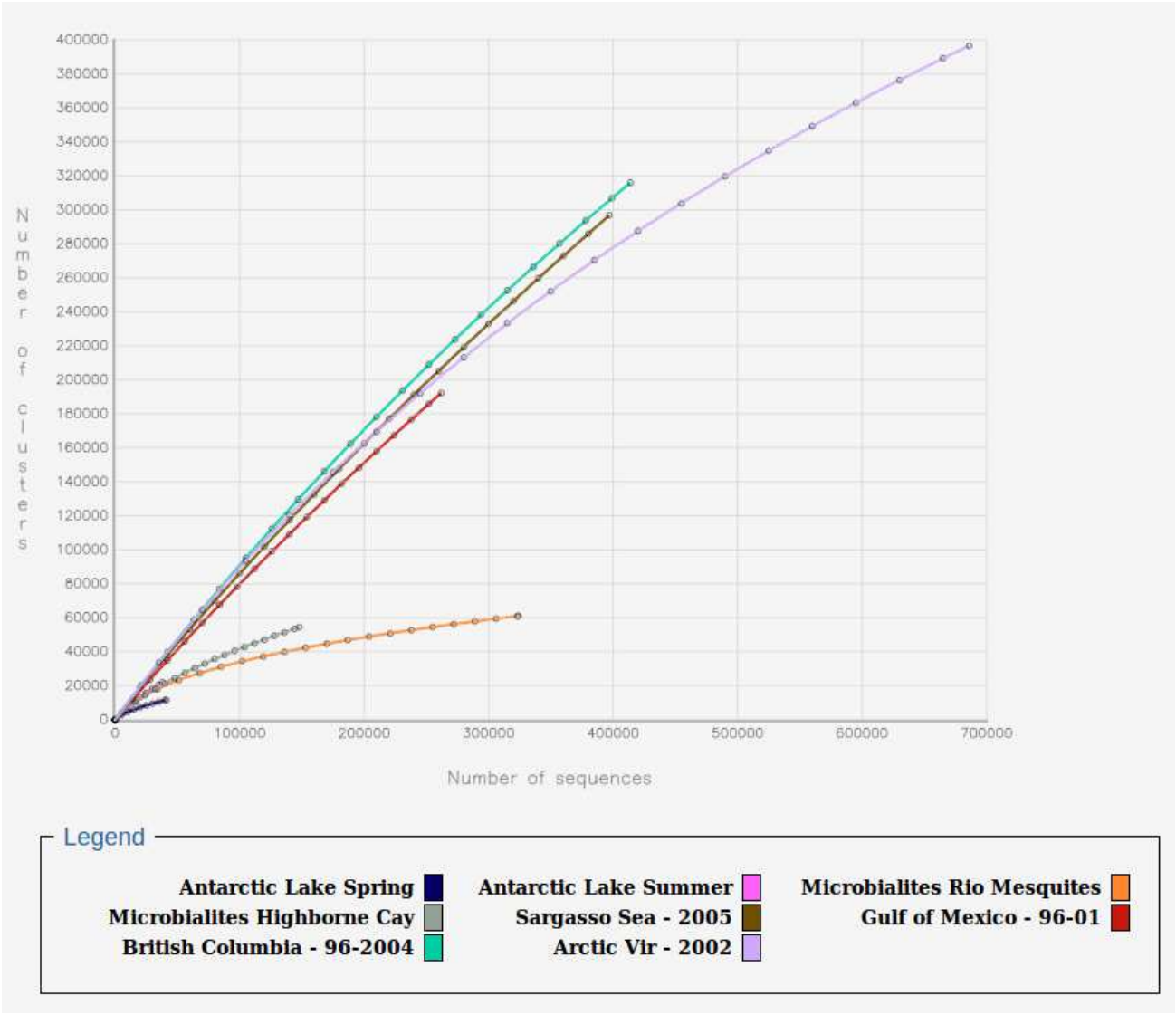
Supplementary Figure 1. Taxonomic composition of the Sargasso Sea virome, obtained with the GAAS tool through Metavir portal, with a threshold of 10^{-5} on the e-value. For the virome selected here, the main identified viruses for this virome are *Microviridae* (*Chlamydia* phage Chp1, Chp2, 3, 4, CPAR39, *Bdellovibrio* phage phiMH-2K), and a *Podoviridae* (*Prochlorococcus* phage P-SSP7). This picture and the text files associated can be downloaded from the web site (in svg format and text format).



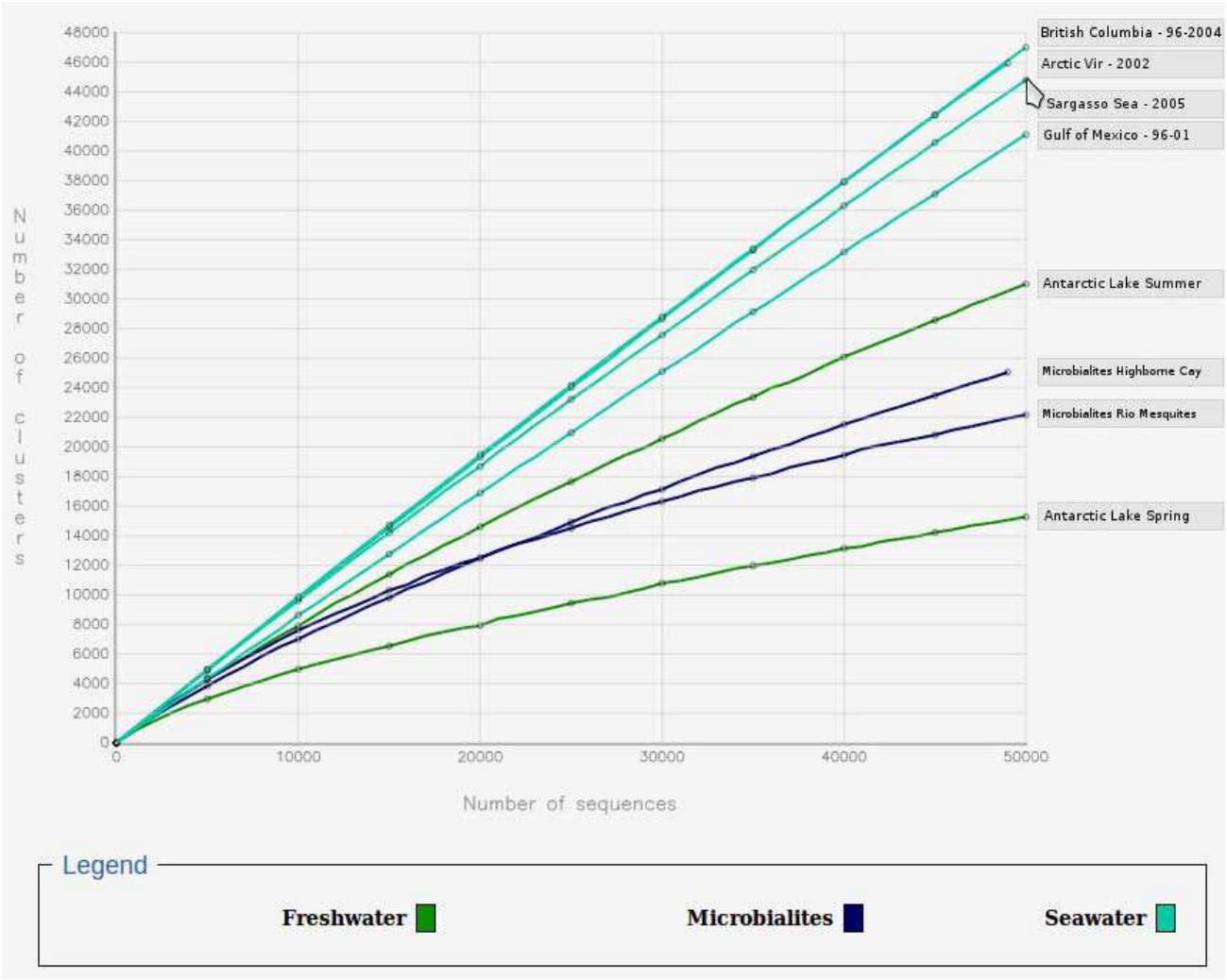
Supplementary Figure 2. Phylogenetic tree drawn for the Sargasso Sea virome using VP1, the marker gene of the Microviridae (major capsid gene present in all the known *Microviridae*). Indeed, the *Microviridae* appeared to be dominant for this particular virome (Fig. S1) . This tree includes the reference sequences of VP1 (in black) plus sequences assembled from the Sargasso Sea Virome (in red), chosen by the user. *Microviridae* from Sargasso Sea appear to be close to other uncultured *Microviridae*, and related to *Chlamydia* phages. For this phylogenetic tree section, a package can be downloaded on Metavir for every association between a marker gene and a virome. This package contains the protein sequences of the virome homologous to the marker in a fasta file, the phylogenetic trees created as text files (newick format) and figures (eps format).



Supplementary Figure 3. Clustering tree based on sequence similarity compiling different aquatic viromes available on Metavir, including the Sargasso Sea virome. The Sargasso Sea virome (at bottom) is clustered with the marine viromes in a group separated from the viromes from freshwater ecosystems and microbialites. This tree was downloaded from the web site as a picture (in eps format, png and pdf being also available). The similarity matrix used to generate the cluster tree is also available in csv format.



Supplementary Figure 4. Rarefaction curves drawn from whole viromes. Here, all the sequences of each virome are clustered into groups using a 75% threshold for percent identity. Curves from different aquatic viromes available via Metavir, including the Sargasso Sea virome (in brown), are reproduced in different colors for each. This is a screen shot of the rarefaction curve figure displayed in Metavir. We provide the matrix used to draw the curve (csv format), matrix that can be easily used in any spreadsheet program to draw the curves.



Supplementary Figure 5. Rarefaction curves drawn from sampled virome. Here, to compare viromes, 50,000 sequences of 100 bp of each virome are clustered into groups using a 75% threshold for percent identity. Curves from different aquatic viromes available via Metavir, including the Sargasso Sea virome, are reproduced in different colors for different types of viromes selected. The gene richness of the Sargasso Sea viral community appears to be very similar to the richness of other marine viromes. On the web site, the name of each virome can be brought up on by hovering the mouse over the dots on the curves. Here, names have been added manually on the right hand side of the figure.

Metavir 2 : analyses comparatives et traitement de séquences assemblées

Plusieurs modifications ont été effectuées dans les mois ayant suivi la mise en ligne de Metavir en septembre 2011. Grâce notamment aux interactions avec les différents utilisateurs, de nouvelles fonctionnalités ont été ajoutées visant à améliorer le confort d'utilisation et à proposer une gamme plus large d'analyses.

Développement des analyses comparatives

Une nouvelle comparaison de viromes a tout d'abord été ajoutée, basée sur la fréquence de mots de taille 2, 3 et 4 au sein des séquences nucléotidiques, comme présenté par Willner et collaborateurs (Willner *et al.*, 2009b). Cet ajout s'inscrit dans l'idée de présenter également sur Metavir des outils existant mais non disponibles sur d'autres serveurs. Toutefois, les résultats obtenus par la comparaison des compositions en nucléotides sont parfois compliqués à analyser, notamment lors de l'étude de viromes très différents. Un nouveau mode de visualisation des comparaisons a également été ajouté, avec la possibilité de remplacer la représentation en dendrogramme par un positionnement multidimensionnel non-métrique (Oksanen *et al.*, 2008), qui permet de placer les différents jeux de données sur un graphique en deux dimensions au sein duquel la distance entre les points est la plus proche possible de la distance génétique estimée entre les jeux de données.

Au niveau de l'analyse des séquences affiliées, différentes compositions taxonomiques sont maintenant proposées basées sur différents seuils, et avec l'utilisation ou non du logiciel GAAS. Ces différentes compositions sont présentées au sein d'un graphique dynamique grâce à l'outil Krona (Ondov *et al.*, 2011). De plus, une comparaison des affiliations taxonomiques a été implémentée, à la fois *via* une série de graphiques au sein de Krona, et sous la forme de tableaux croisés ("*heatmap*") interactifs générés à la volée. Cette vision comparée des fractions affiliées est ainsi un complément à la comparaison de viromes complets.

Afin de mieux caractériser ces similarités observées entres les séquences de différents viromes et les génomes complets, un outil de graphiques de recrutement a été développé, positionnant le long d'un génome de référence les hits détectés par BLAST pour un ou plusieurs viromes. Il est ainsi possible de visualiser l'ensemble des séquences similaires à un génome donné pour une série de viromes, soit en fonction de leur pourcentage de similarité au génome de référence (chaque séquence est alors représentée par un point), soit sous la forme

d'une distribution le long du génome (le nombre de séquences par région de 500 nucléotides est alors indiqué sous la forme d'un histogramme).

Enfin, la partie phylogénie a elle aussi été légèrement modifiée, à la fois sur le fond de l'outil et dans la présentation des résultats. Tout d'abord, la procédure a été adaptée afin de permettre la génération d'arbres à partir de plusieurs viromes, ce qui permet de comparer les séquences de viromes entre elles au sein de phylogénies, et potentiellement de déterminer de nouveaux clades. Afin de pouvoir visualiser correctement ces arbres, devenus de plus en plus complexes, un Plug-in JavaScript dédié (JsPhyloSVG ; (Smits & Ouverney, 2010)) a été adapté afin de présenter les arbres de manière interactive et automatiquement annotés (mise en avant des séquences environnementales, identification des groupes de référence). De nouveaux marqueurs ont également été ajoutés pour étendre le spectre des groupes susceptibles d'être étudiés, notamment aux virus à ARN.

Annotation de fragments génomiques assemblés

Si un certain nombre de modifications a été apporté aux outils proposés dans la première version de Metavir, ces derniers sont adaptés aux séquences courtes et non assemblées, comme celles constituant les premiers viromes publiés. Or, l'évolution très rapide des techniques de séquençage à haut-débit a profondément modifié le type de données

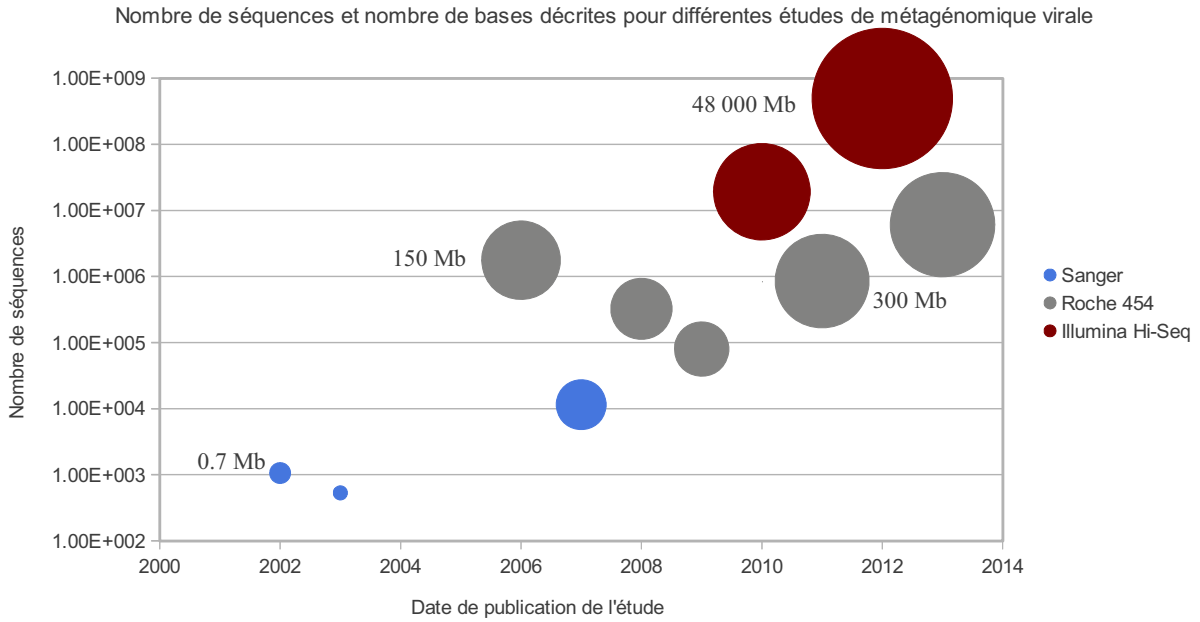


Figure I.1 : Évolution du nombre de nucléotides et de séquences analysés dans différentes études de métagénomiques virales publiées depuis 2003*. Pour chaque étude, le nombre de séquences étudiées (en ordonnée) et la quantité de paire de bases (surface du cercle) sont reportés en fonction de l'année de publication.*

générées. En effet, si les premiers jeux de données étaient composés d'un millier de séquences de quelques centaines de paires de bases, les premiers viromes associés aux nouvelles techniques de séquençage (principalement les technologies Roche 454) étaient formés de centaines de milliers de séquences, d'abord d'une centaine de paires de bases, puis de plus en plus longues, qui pouvaient être directement analysées. L'apparition des séquenceurs de type HiSeq (Illumina) a fait franchir une nouvelle étape aux analyses métagénomiques, en proposant cette fois plusieurs dizaines de millions de séquences, généralement d'une centaine de paires de bases (Figure I.1). Il est ainsi quasiment impossible d'analyser indépendamment ces séquences brutes, et surtout beaucoup plus intéressant et informatif de passer par un assemblage, puisqu'un tel nombre de séquence permet d'obtenir une couverture suffisante pour l'assemblage de longs fragments génomiques, voire de génomes complets. Un ensemble d'outils d'analyse a donc été mis en place sur le serveur Metavir pour traiter au mieux ces séquences assemblées.

A l'image des analyses effectuées sur les séquences brutes, une affiliation taxonomique est réalisée pour les gènes prédits à partir d'une comparaison par BLAST avec la base de données RefseqVirus du NCBI. En complément de cette comparaison aux génomes viraux complets, les protéines prédites sont comparées aux domaines protéiques de la base de donnée PFAM *via* le logiciel hmmscan (Eddy, 2011). Basé sur l'utilisation de profils de Markov, cet outil permet de comparer dans un temps raisonnable les séquences métagénomiques à une base de donnée généraliste, et ainsi déterminer une fonction potentielle pour une protéine prédite même en l'absence de similarité avec une séquence virale.

Visualisation des fragments génomiques annotés

Au-delà de la composition de la communauté, l'analyse de contigs nécessite la visualisation individuelle des séquences assemblées. En effet, chaque contig peut être porteur d'information unique, il est donc nécessaire de compléter la vision quantitative des affiliations par une vision qualitative de chaque élément. Ainsi, une fiche au format GenBank est automatiquement générée pour chaque contig, synthétisant l'ensemble des annotations réalisées par Metavir : circularité du contig, coordonnées des gènes prédits, affiliation taxonomique du contig, affiliation taxonomique et fonctionnelle de chaque gène. Ce type de fichier peut ensuite être utilisé dans la plupart des outils d'annotation de génomes, qui sont généralement utilisés pour traiter ces séquences assemblées. Une visualisation en ligne est également proposée, avec la génération de cartes interactives permettant d'observer en une figure à la fois la structure du contig et les informations individuelles de chaque gène. Un soin

particulier a donc été apporté à cet aspect de présentation des résultats, afin qu'un utilisateur puisse naviguer au mieux au sein des milliers de séquences assemblées à sa disposition, et isoler et identifier le plus facilement possible celles qui présentent le plus d'intérêt pour son étude. A cet effet, un système de filtre a été conçu afin que l'utilisateur puisse sélectionner les contigs sur leur taille, leur identifiant, leur affiliation taxonomique, ou les affiliations taxonomiques et fonctionnelles des gènes qu'il contient. Cette interface permet ainsi à l'utilisateur de pouvoir limiter le nombre de séquences à explorer et à analyser, et optimise l'utilisation des ressources du serveur en limitant la taille des requêtes.

Il est également indispensable de comparer les fragments génomiques entre eux et avec les génomes de référence, à la fois au niveau des gènes (niveaux de similarité, domaines conservés) mais aussi du point de vue de l'organisation de ces derniers le long des génomes. De plus, de par l'organisation en modules de gènes des génomes viraux (en particulier des phages), plusieurs génomes de référence peuvent être associés à un même contig, et il est alors nécessaire d'identifier précisément les différentes régions similaires aux génomes de référence. Afin de faciliter ce travail, deux outils ont été associés aux cartes de contigs sur le serveur Metavir : un réseau de contigs et génomes, et une comparaison de cartes. L'utilisation de la représentation en réseau grâce au plug-in Cytoscape-web (Lopes *et al.*, 2010) permet de visualiser les similarités entre les gènes de différents contigs (au sein d'un virome ou entre viromes) ainsi que les liens entre contigs et génomes de référence. Il est donc possible de positionner un contig particulier par rapport aux génomes connus, d'identifier des contigs partageant un gène ou un ensemble de gènes, et de repérer des événements de transferts horizontaux ou de recombinaisons. En complément de cette visualisation en réseau, un utilisateur peut sélectionner un ensemble de séquences d'intérêt (contigs ou génomes), et générer pour cette sélection un graphique comparant les cartes deux à deux.

Enfin, les outils conçus pour les séquences brutes de viromes ont été adaptés au mieux aux contigs. Les courbes de raréfaction ont ainsi été complétées par une clusterisation des séquences protéiques prédites, plus informative que la clusterisation des contigs complets. Les graphiques de recrutement ont été remplacés par un histogramme présentant le nombre de contigs similaires pour chaque gène d'un génome de référence. La procédure de génération d'arbres phylogénétiques a également été adaptée pour se baser sur les gènes prédits, et les comparaisons globales de viromes ont été conservées.

Article II

METAVIR 2: virome comparative analysis and annotation of assembled genomic fragments.

Simon Roux[1,2*], Jérémy Tournayre[1,2], Antoine Mahul[3], Didier Debroas[1,2] and François Enault[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand

[2]CNRS, UMR 6023, LMGE, F-63177 Aubière

[3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France

Soumis à **BMC Bioinformatics**

Metavir 2 : virome comparative analysis and annotation of assembled genomic fragments.

Simon Roux[1,2], Jeremy Tournayre[1,2], Antoine Mahul[3], Didier Debroas[1,2], François Enault[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, Clermont-Ferrand

[2]CNRS, UMR 6023, LMGE, Aubiere

[3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand

Keywords: viral metagenomics, phage genomics, web-server

Abstract

Background. Viruses are the most abundant and diverse biological entities on Earth, and metagenomics is a well-fitted and powerful tool to explore viral communities. The field of viral metagenomics is young but quickly evolving, mainly due to the fast development and cost reduction of high-throughput sequencing (HTS) technologies. Indeed, these technological developments led to a quantitative modification of virome data, with a sharp increase in the number of viromes published these last years, as well as a qualitative change linked to the considerable amount of sequence data now generated, which makes it possible to assemble large genome fragments. Thus, these technological changes demand a same-paced development and implementation of adapted bioinformatics tools.

Results. To improve the analysis of multiple datasets, all Metavir tools have been adapted to support comparative analysis of viromes. In addition to the sequence comparison previously provided, viromes can now be compared through their k-mer frequencies, their taxonomic composition (through interactive piecharts or heatmaps), recruitment plots involving multiple viromes, and phylogenetic trees including here again sequences from different datasets. To handle assembled datasets, a brand new section has been specifically designed and implemented. This section includes an annotation pipeline for viral contigs (gene prediction, similarity search against reference viral genomes and protein domains) and an extensive comparison between annotated contigs and reference genomes. Contigs and their annotations can be explored on the website through specifically developed dynamic genomic maps and interactive networks.

Conclusion. Overall, the new features of Metavir 2 allow users to explore and analyze viromes composed of raw reads or assembled fragments through a set of adapted tools and a user-friendly interface.

Background

Viruses are the most abundant biological entities in the biosphere [1]. At first seen mostly as infectious agents [2, 3], viruses are now considered as major players in natural ecosystems and their associated cycles and balances [4, 5]. Viral communities are known to be mostly composed of uncharacterized strains, which genomes encode genes both mostly uncharacterized and at the same time highly diversified [6, 7]. However, the description and characterization of new viral genomes is made difficult by the unsuitability of several techniques usually applied in molecular ecology. First, most of the micro-organisms are still impossible to cultivate in the lab for now, hence preventing the culture and isolation of their associated viruses. Second, sequencing of a universal marker gene after a PCR amplification, as it is currently done for micro-organisms through the analysis of ribosomal RNA, can not be extended to viruses as no universal gene is available.

Metagenomic approaches (*i.e.* random sequencing of the genetic pool isolated from natural samples) circumvent these limitations, as no prior knowledge on the targeted sequences is required, and are therefore increasingly applied to viral communities. Protocols are now well established to extract and isolate the encapsidated fraction [8–10], and viral metagenomes (viromes) have now been generated from almost all types of biomes and samples. Beyond the description and characterization of the viral genomic diversity, viromes are useful toward more general questions such as biogeography and dispersion of viral particles [11, 12], evolution and origin of viruses [13] or epidemiology [14].

Recent advances in next-generation sequencing and in sequence assembly techniques (see for example [15–17]) led viral metagenomics a step further, by providing access to large genomic fragments rather than only short reads. Indeed, small complete viral genomes can be assembled from 454-generated viromes [18–21] and the use of Illumina Hi-Seq 2000 sequencing made it possible to assemble complete genomes of up to 60 Kb [22–24]. These long genomic sequences are of great value while studying environmental viral genomes and offer the possibility to gain unique insights into the viral families retrieved. Indeed, the genomic content and architecture of families of interest provide new ways of understanding their ecological functioning and evolution.

Two web-servers are currently available for a comprehensive virome analysis : Metavir [25], and Virome [26]. A pipeline (the Viral Metagenome Affiliation Pipeline [27]) was also described but to our knowledge is not available neither as a stand alone software or through a web page. None of these bioinformatic tools were designed for the analysis of assembled datasets and the absence of adapted tools for such assembled viromes was pinpointed as a major bottleneck for viral metagenomics studies [3, 26]. Moreover, the growing number of viromes publicly available calls for the development of comparison strategies to go beyond a separate analysis of each dataset. Here, we introduce a new version of Metavir that tackles these two limitations. Metavir 2 includes (i) new ways to compare datasets, notably a new tool to accurately describe similarities between sequences of one or several viromes and reference genomes, and (ii) a whole new section which forms the first set of tools designed for a comprehensive analysis of assembled virome sequences.

Implementation

Input and metadata

Registered users can upload their own sequence datasets, either short reads or assembled contigs, in a private space. Input data are checked for being only composed of DNA sequences in fasta format (compressed files in zip, gzip or tar.gz format are accepted). Due to the size of Illumina's raw datasets (several tens of Gb) and computing time required for assembling each dataset, the assembly step can not be computed through Metavir. Furthermore, a wide range of softwares are available for this step and the choice depends on the type of the sequencing and the nature of the sample : Newbler (454 Life Sciences) is the main software used so far for 454 data [21, 28, 29], and Illumina data can be assembled with Idba_ud [16], SOAP [30], MetaVelvet [31] or OptiDBA [17].

A set of public viromes is also already available for users to compare with their dataset(s). These viromes are sorted into projects, and linked to the manuscript describing their analysis when available. Metadata such as the type of sample from which the virome was sequenced, the location, depth, and temperature of sampling point, and the sequencing technology used to generate the dataset, can be added.

Section 1 : Tools to analyze raw datasets (unassembled reads)

Taxonomic composition

Virome reads are first compared to the complete viral genomes of the RefSeq Virus database using BLAST. The taxonomic composition is then determined using either raw best hit numbers or best hit numbers normalized by genome length using the GAAS tool [32]. Krona, a tool dedicated to the interactive visualization of hierarchical data in a Web browser [33], is now used to generate dynamic chart representing taxonomic composition of one or more viromes. A custom-designed javascript program has also been implemented to visualize compositions as interactive heatmaps, with each column representing a dataset and each row a group of viral species. Columns can be switched by mouse drag and drop. Viral species are classified according to the up-to-date NCBI taxonomy, and viral groups can be folded and unfolded with a mouse click.

k-mer frequency bias

A virome comparison based on k-mer nucleotide frequencies (di- tri- and tetra-nucleotides are available) has been implemented as described by Willner and collaborators [34]. Unlike the other available comparison method, based on sequence similarity (generated using reciprocal tBLASTx) and requiring datasets containing at least 50,000 sequences of 100bp, k-mer nucleotide frequencies can be computed for all datasets without size restriction. Briefly, k-mer frequency distribution bias are computed by a custom Perl script and then compared for each pair of viromes. Pairwise distances between viromes are stored in a matrix, which can be used as input either in a hierarchical clustering or a non-metric multidimensional scaling. Both analysis are computed with R [35] with pvclust [36] and vegan [37] libraries respectively. The non-metric multidimensional scaling (NMDS) is now also available for virome comparison based on sequence similarities, available in Metavir 1.

Phylogenetic analyses

To speed up the phylogenetic pipeline without extensive modifications of the results, the phylogenetic pipeline of Metavir now computes phylogenetic trees with FastTree [38]. The visualization of the resulting trees was also modified: online phylogenetic tree are now dynamic, thanks to an adaptation of the jsPhyloSVG javascript plugin [39]. Tree displays can be circular or linear, subtrees can be merged, and informations on the origin and affiliation of the sequence of each node can be obtained by clicking on the associated leaf.

Individual viral genome recruitment plots

Using the best BLAST hit results against Refseq Virus, each virome sequence is affiliated to a unique viral genome. For any selected viral genome, two types of recruitment plots are then available : (i) a scatter plot displaying each hit read as a dot depending on the position on the genome (on the x-axis) and the identity percentage of the BLAST hit (on the y-axis), and (ii) an histogram presenting the number of hits for each 500-nt long genome part. These plots are generated using the ggplot2 R library [40]. Once the first plot is displayed, viromes that contain sequences recruited by the selected genomes are listed and can be added on the same plot. When several datasets are selected, a color is attributed to each virome, used to color dots (in scatter plots) or stacked histograms (in histograms).

Section 2 : Assembled viromes annotation and display

Contig annotation

Open reading frames (ORFs) are first predicted for each contig through MetaGeneAnnotator [41]. A custom Perl script was designed to detect circular contigs by looking for identical k-mer at the two ends of the sequences. Each circular contig is then trimmed to remove all redundant parts. In order to be able to predict genes spanning the origin of circular contigs, a temporary version of circular contigs is used in the ORF prediction software, in which the first 1,000 nucleotides are duplicated and added at the contig's end. It has to be noted that this detection of circular contigs will not be effective for contigs computed with assembler like Newbler which already detect and remove such similarity between contig ends.

All predicted ORFs are then compared to several databases, namely the Refseq Virus protein database from the NCBI using BLASTp [42], with a threshold of 10^{-3} on e-value, and the PFAM database of protein domains (version 26.0 ; [43]) using HMMscan [44], with a threshold of 30 on score. A direct comparison of ORFs within a virome is also computed through a BLASTp with the same threshold of 10^{-3} on e-value.

The taxonomic composition and sequence diversity measured through clustering are also different for datasets made of long genomic sequences. Using the BLASTp results against reference viruses, three types of taxonomic compositions are computed for each dataset. These compositions are based on (i) best BLAST hit affiliation of each predicted gene, (ii) best BLAST hit affiliation of each contig, and (iii) lowest common ancestor affiliation of each contig. This LCA affiliation is designed to take into account the multiple hits on a single contig : up to five affiliated genes (if available) are considered for each contig, and the affiliation is made at the highest common taxonomy level of the best BLAST hit from these selected genes.

Finally, different clusterings of the predicted ORFs are computed. A global protein sequence clustering with three different thresholds (75, 90 and 98% of similarity) is performed using Uclust [45]. Another clustering is based on protein sequences domain alignments : ORFs are first ordered by size, and used iteratively as a seed for a jackhmmer search [44]. All ORFs recruited by the seed are gathered in a cluster with this seed, and removed from further iterations. Once computed, the domain-based ORFs clusters are affiliated to one or more PFAM domain based on the affiliation of their members. These clusterings are displayed through the rarefaction curve tool.

Contig display

When an assembled virome is selected, a new “contig maps” page now provides general informations about ORFs prediction and contig affiliations, as well as a filter used to reduce the number of contigs to use in subsequent displays (contig maps and networks). Jquery is used to display this interactive filter, allowing to select contigs based on taxonomic or functional affiliation of a predicted gene, and contig size, name or taxonomic affiliation.

A summary of contig annotations is also available for each individual contig as an interactive map, drawn using RaphaelSVG and the Raphael-zpd plugin. Each gene affiliation to Refseq viral genomes and PFAM protein domains is indicated when available, and genes can manually be further investigated as nucleotide and protein sequences are

displayed by clicking on the gene either on the map or on the gene table below. The same contig annotations can be downloaded as csv tables, summarized by contig or detailed for each ORFs.

Similarities between contigs and viral genomes and between different contigs can be displayed within an interactive network. In order to take into account all relevant similarities between contigs and genomes, and not only the best BLAST hit, all BLAST hits against Refseq database fulfilling the criteria for taxonomic affiliation (i.e. e-value lower than 10^{-3}), and having a bit-score within a 10% margin from the best BLAST hit bit-score for this ORF are displayed in the contig network. Cytoscape-web [46] is used to display this network, allowing users to see contigs and reference genomes as nodes, and sequence similarities as edges. Different options are available to customize the network, such as the coloring of edges based on BLAST bit-score, the display of only one edge between two similar contigs or of one edge for each ORFs similarity, or the coloring of genome nodes based on the taxonomy. Another set of filters is also proposed to reduce the number of nodes or edges displayed on screen.

Associated with this network, a contig map comparison tool can be used to display collinearity between contigs and genomes or other contigs selected on the network. This comparisons are displayed through RaphaelSVG and Raphael-zpd. Name and affiliation of each gene is displayed when clicked, and a Jquery pop-up is used to change the sequence order within the plot.

Common framework

Automatic database update

As the RefseqVirus database is quickly growing (80 new genomes are added on average with every bimonthly release), each new release is automatically downloaded and used as the new reference database. Taxonomic composition, gene affiliation (for contig dataset), and recruitment plots of public projects are automatically updated with each release, whereas the update of private projects must be manually required by the user through a button on the taxonomy page.

Results and graphics download

All sequence datasets used in a Metavir analysis are available for download in fasta format (affiliated and uncharacterized sequences, sequences included in phylogenetic trees and sequences included in recruitment plots). All tables from the taxonomic heatmap or summarizing the contig ORF affiliations are downloadable as csv files, which can be imported in all current spreadsheets. For each recruitment plot generated, the list of each gene with the associated number of hits (for each virome) can be downloaded as a csv file. Alternatively, the sequences used to generate the plots (i.e. virome reads with a best BLAST hit against the selected genome) can be downloaded either in a single fasta file, or in a different file for each gene.

Contig annotations are available in GenBank file format, which can be used in many downstream tools like Artemis [47] or Easyfig [48]. These GenBank files contain the lowest common ancestor affiliation of the contig, as well as the best BLAST hit affiliation of each ORF, the functional annotation of each ORF in PFAM domain, and the sequences associated with each predicted CDS.

All interactive charts and pictures, like contig maps, contig comparisons or phylogenetic trees can be downloaded in svg format, a publication-ready vectorial format easy to modify using graphics softwares. Static charts generated with R are available to download in pdf and png file format.

Finally, the contig networks can be downloaded in a set of different formats, including graphml and xgmml, ready to be imported in the desktop version of Cytoscape for further analyses and annotations (for an optimal import of Metavir-generated networks, Cytoscape 3 is required).

Case study: using Metavir to analyze the human gut virome

Two different datasets from the human gut virome were chosen to illustrate the results that can be obtained with Metavir 2. First, a set of 16 metagenomic libraries from virome samples was used to illustrate the section dedicated to unassembled datasets ([49] ; project "Human Gut Diet" on Metavir). These metagenomes, sequenced with 454 GS Titanium (884628 reads of 350 bp / 310 Mb), were initially designed to study the dynamics of human gut viral community during a perturbation by a dietary intervention. Two individuals were fed a high fat/low fiber diet (H1 and H2), three were fed a low fat/high fiber (L1, L2 and L3) and one was on an ad-lib diet (X). Samples were collected at up to four time points (days 1, 2, 7 and 8). The second dataset is an assembly of Illumina Hi-Seq 2000 reads (5.6 Gb of 100 bp reads from 6 samples) obtained by sequencing virome samples of healthy individuals ([17] ; virome "Human gut – All subjects" from project "Human Gut Assembly" on Metavir). This assembled dataset was used here to illustrate the possibilities offered by the new section dedicated to the analysis of contigs.

Results and Discussion

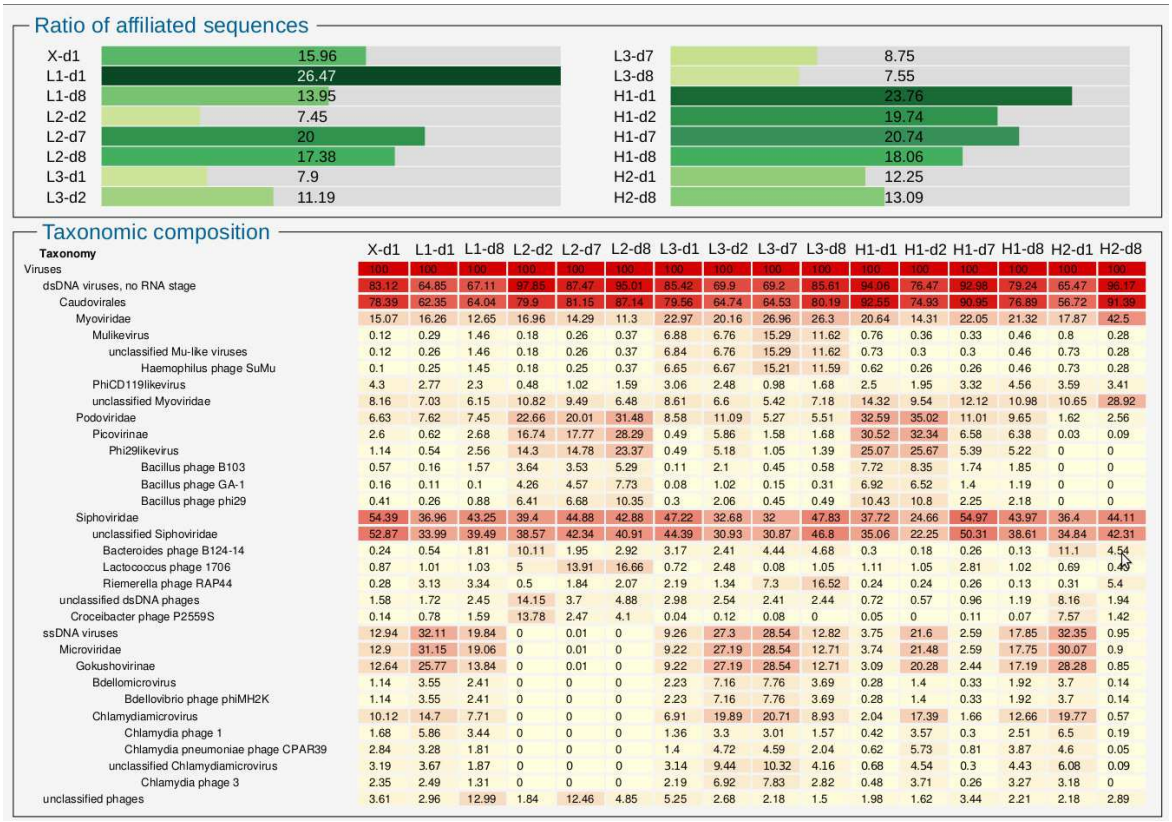
Metavir now provides a single web interface to analyze the two existing types of datasets that can be uploaded by registered users : (i) viromes composed of raw reads, mostly generated using pyrosequencing technology and (ii) viromes assembled into contigs, a strategy either possible with datasets sequenced with Illumina technology or 454. The different novelties introduced by this manuscript will be illustrated using both types of datasets (unassembled 454 reads [49] and Illumina assembled contigs [17]), all from human gut samples.

Additions to the unassembled read section

The initial version of Metavir [25] was only dedicated to the analysis of unassembled read datasets and was made of a set of standard and specifically designed tools : (i) a taxonomic composition assessment, (ii) the generation of rarefaction curves, (iii) the comparison of datasets using pairwise virome sequence similarity and (iv) a phylogenetic tree pipeline. Most viral metagenomes published still consist of unassembled reads, and several samples are often sequenced in a single 454 pyrosequencing run that can be easily multiplexed. Multiple datasets allow to study spatial or temporal dynamics in environmental communities [11, 23, 50–52] or different individuals subjected to different conditions for eukaryote-associated viromes (*e.g.* different diets in [49]). In this context, the comparison of multiple datasets was our major focus while extending this section of Metavir. Indeed, in the initial version of Metavir, rarefaction curves and reciprocal tBLASTx comparison were the only tools allowing to compare different datasets. In this new version, all tools can now be used to compare viromes. Furthermore, most of these tools were improved with a special attention on the display of results and the user experience. A brand new tool was also added: the recruitment plot analysis, which makes it possible to accurately study the similarities between virome reads and a viral genome of interest.

Taxonomic composition

Individual taxonomic compositions for each virome are based on the comparison of virome reads with complete known viral genomes, using either raw results or results normalized by genome length [32], and displayed through interactive chart prepared by Krona [33]. Furthermore, virome compositions can now be visually compared in two ways : multiple compositions can be merged on the same Krona chart, or they can be displayed in a more hierarchical way through an in-house developed interactive heatmap. As an example, such a taxonomic heatmap was generated for the 16 datasets from the human gut (Figure 1). This heatmap allows to quickly visualize that these datasets only exhibited similarities with bacteriophages, in accordance with the results presented in Minot *et al.* ([49], figure 2c). Even if the same bacteriophage groups are found in the different datasets, the proportion of these groups differ between each virome: *Myoviridae* constitute between 11 and 42% of the



Taxonomy	X-d1	L1-d1	L1-d8	L2-d2	L2-d7	L2-d8	L3-d1	L3-d2	L3-d7	L3-d8	H1-d1	H1-d2	H1-d7	H1-d8	H2-d1	H2-d8
Viruses	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100	100
dsDNA viruses, no RNA stage	83.12	64.85	67.11	97.85	87.47	95.01	85.42	69.9	69.2	85.61	94.06	76.47	92.98	79.24	65.47	96.17
Caudovirales	78.39	62.35	64.04	79.9	81.15	87.14	79.56	64.74	64.53	80.19	92.55	74.93	90.95	76.89	56.72	91.39
Myoviridae	15.07	16.26	12.65	16.96	14.29	11.3	22.97	20.16	26.96	26.3	20.64	14.31	22.05	21.32	17.87	42.5
Mulikevirus	0.12	0.29	1.46	0.18	0.26	0.37	6.88	6.76	15.29	11.62	0.76	0.36	0.33	0.46	0.8	0.28
unclassified Mu-like viruses	0.12	0.26	1.46	0.18	0.26	0.37	6.84	6.76	15.29	11.62	0.73	0.3	0.3	0.46	0.73	0.28
Haemophilus phage SuMu	0.1	0.25	1.45	0.18	0.25	0.37	6.65	6.67	15.21	11.59	0.62	0.26	0.26	0.46	0.73	0.28
PhiCD119likevirus	4.3	2.77	2.3	0.48	1.02	1.59	3.06	2.48	0.98	1.68	2.5	1.95	3.32	4.56	3.59	3.41
unclassified Myoviridae	8.16	7.03	6.15	10.82	9.49	6.48	8.61	6.6	5.42	7.18	14.32	9.54	12.12	10.98	10.65	28.92
Podoviridae	6.63	7.62	7.45	22.66	20.01	31.48	8.58	11.09	5.27	5.51	32.59	35.02	11.01	9.65	1.62	2.56
Picovirinae	2.6	0.62	2.68	16.74	17.77	28.29	0.49	5.86	1.58	1.68	30.52	32.34	6.58	6.38	0.03	0.09
Phi29likevirus	1.14	0.54	2.56	14.3	14.78	23.37	0.49	5.18	1.05	1.39	25.07	25.67	5.39	5.22	0	0
Bacillus phage B103	0.57	0.16	1.57	3.64	3.53	5.29	0.11	2.1	0.45	0.58	7.72	8.35	1.74	1.85	0	0
Bacillus phage GA-1	0.16	0.11	0.1	4.26	4.57	7.73	0.08	1.02	0.15	0.31	6.92	6.52	1.4	1.19	0	0
Bacillus phage phi29	0.41	0.26	0.88	6.41	6.68	10.35	0.3	2.06	0.45	0.49	10.43	10.8	2.25	2.18	0	0
Siphoviridae	54.39	36.96	43.25	39.4	44.88	42.88	47.22	32.68	32	47.83	37.72	24.66	54.97	43.97	36.4	44.11
unclassified Siphoviridae	52.87	33.99	39.49	38.57	42.34	40.91	44.39	30.93	30.87	46.8	35.06	22.25	50.31	38.61	34.84	42.31
Bacteroides phage B124-14	0.24	0.54	1.81	10.11	1.95	2.92	3.17	2.41	4.44	4.68	0.3	0.18	0.26	0.13	11.1	4.54
Lactococcus phage 1706	0.87	1.01	1.03	5	13.91	16.66	0.72	2.48	0.08	1.05	1.11	1.05	2.81	1.02	0.69	0.43
Riemerella phage RAP44	0.28	3.13	3.34	0.5	1.84	2.07	2.19	1.34	7.3	16.52	0.24	0.24	0.26	0.13	0.31	5.4
unclassified dsDNA phages	1.58	1.72	2.45	14.15	3.7	4.88	2.98	2.54	2.41	2.44	0.72	0.57	0.96	1.19	8.16	1.94
Croceibacter phage P2559S	0.14	0.78	1.59	13.78	2.47	4.1	0.04	0.12	0.08	0	0.05	0	0.11	0.07	7.57	1.42
ssDNA viruses	12.94	32.11	19.84	0	0.01	0	9.26	27.3	28.54	12.82	3.75	21.6	2.59	17.85	32.35	0.95
Microviridae	12.9	31.15	19.06	0	0.01	0	9.22	27.19	28.54	12.71	3.74	21.48	2.59	17.75	30.07	0.9
Gokushovirinae	12.64	25.77	13.84	0	0.01	0	9.22	27.19	28.54	12.71	3.09	20.28	2.44	17.19	28.28	0.85
Bdellomicrovirus	1.14	3.55	2.41	0	0	0	2.23	7.16	7.76	3.69	0.28	1.4	0.33	1.92	3.7	0.14
Bdellovibrio phage phiMH2K	1.14	3.55	2.41	0	0	0	2.23	7.16	7.76	3.69	0.28	1.4	0.33	1.92	3.7	0.14
Chlamydiamicrovirus	10.12	14.7	7.71	0	0	0	6.91	19.89	20.71	8.93	2.04	17.39	1.66	12.66	19.77	0.57
Chlamydia phage 1	1.68	5.86	3.44	0	0	0	1.36	3.3	3.01	1.57	0.42	3.57	0.3	2.51	6.5	0.19
Chlamydia pneumoniae phage CPAR39	2.84	3.28	1.81	0	0	0	1.4	4.72	4.59	2.04	0.62	5.73	0.81	3.87	4.6	0.05
unclassified Chlamydiamicrovirus	3.19	3.67	1.87	0	0	0	3.14	9.44	10.32	4.16	0.68	4.54	0.3	4.43	6.08	0.09
Chlamydia phage 3	2.35	2.49	1.31	0	0	0	2.19	6.92	7.83	2.82	0.48	3.71	0.26	3.27	3.18	0
unclassified phages	3.61	2.96	12.99	1.84	12.46	4.85	5.25	2.68	2.18	1.5	1.98	1.62	3.44	2.21	2.18	2.89

Figure 1 : Taxonomic composition (best hit ratio) of the 16 datasets from the human gut viromes from Minot et al. (2011). Viral species are classified according to the up-to-date NCBI taxonomy, and groups can be folded and unfolded with a mouse click. Columns have been re-ordered trough mouse drag and drop to gather datasets from each subject. Samples are named according to the diet (X : ad-lib diet, H : high fat/low fiber diet, L: low fat/high fiber) and day of the sample collection after the beginning of the experiment (d1, d2, d7 and d8).

different viromes, *Podoviridae* between 2 and 35%, *Siphoviridae* between 24 and 55% and *Microviridae* between 0 and 31%.

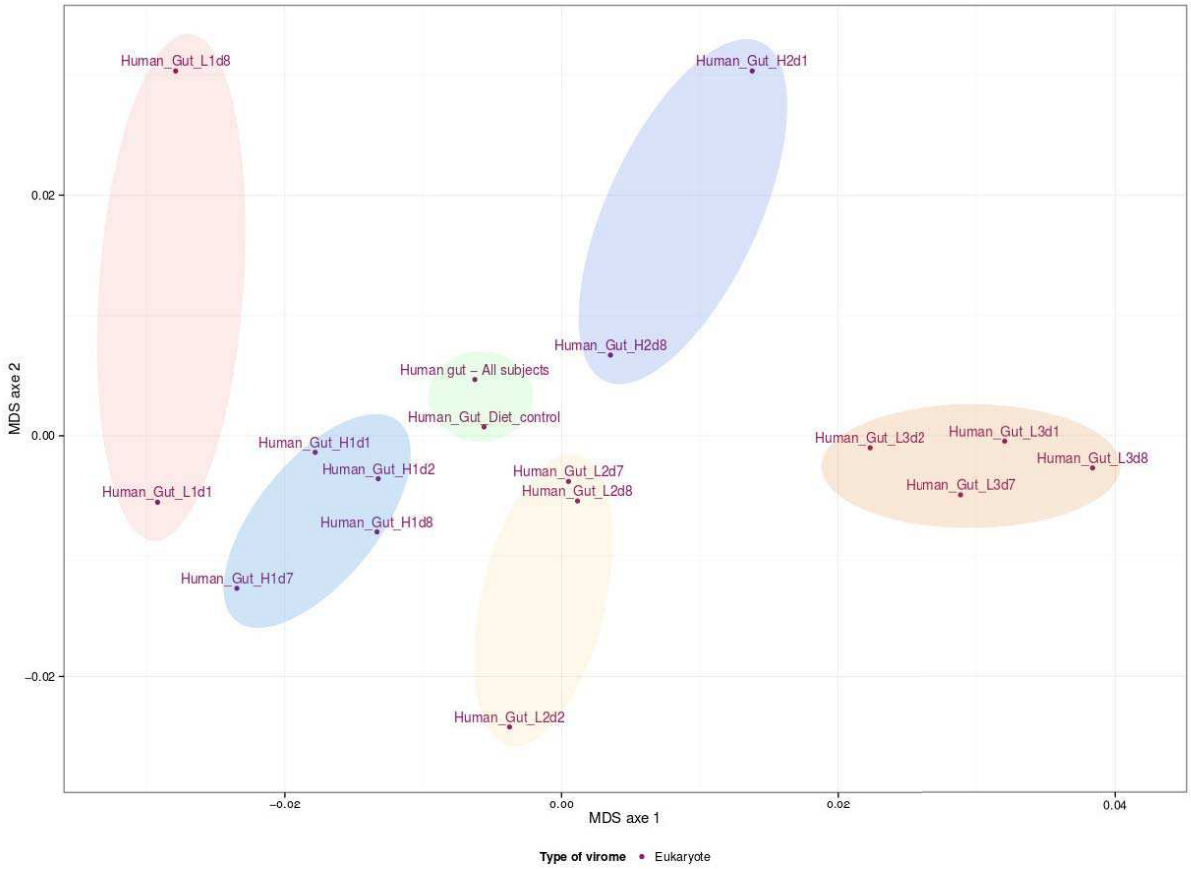
k-mer frequency bias

A recurrent observation in analyses of virome data is that the majority of reads has no similarity to any known viral sequence [53], as can be noted for human gut viromes (top of Figure 1). In this context, methods that consider viromes in their entirety rather than only their small affiliated fraction are of particular interest. K-mer nucleotide frequency bias was proved to be able to distinguish viromes from different biomes and is now available in Metavir. This analysis was applied to the 16 human gut datasets using 4-mer nucleotides (tetranucleotides) and a non-metric multidimensional scaling (Figure 2). Results are here again similar to those obtained in Minot et al. ([49], figure 5A): even though the diet (X, H, L) seems to modify the viral communities, the major factor of differences between datasets is the individuals in themselves as datasets of each subject can be grouped (X1, H1, H2, L1, L2 and L3). Furthermore, viromes from subjects on the same diet do not seems to homogenize over time indicating that each individual contains a unique virome that is globally stable. Thus, this analysis based on all the sequences and not just on the affiliated ones provides informations that cannot be deduced out of the taxonomic compositions.

Phylogenetic analyses

Phylogenetic analysis is of particular interest to study specific viral groups and such analysis was made available in the first version of Metavir [25]. As no universal phylogenetic markers are available for viruses, several marker genes were available for the major viral groups and the list of markers has been expanded to 13 markers, mostly following users' requests. In Metavir 1, reads from a chosen virome detected as homologous to a selected marker were used to compute a tree including both these virome reads and reference sequences. However, the lack of reference strains close to most viruses in most ecosystems limits the interest of such analysis as it often results in the generation of environmental clades far from references. However, samples of similar origin and nature often harbor closely related viruses [7, 12, 52]. Metavir 2 now offers the opportunity to compute phylogenetic trees that include reads from different viromes, which can thus help to gain a better view of the diversity in each sample as well as putative links between samples in a precise manner. As an example, we conducted such an analysis on the *Picovirinae,* a subfamily of *Podoviridae* that is one of the most abundant group in 5 of the 16 human gut viromes, these 5 samples being from two individuals (L2 and H1 ; Figure 1). A typical protein primed DNA polymerase used for replication is conserved inside this family and this gene was used as a marker to determine the phylogenetic relationships of the phages retrieved in these human gut viromes (Figure 3). As expected, all sequences retrieved are gathered near bacteriophages, and no virome reads appear to be linked to either archeal (Salterprovirus) or eukaryotic viruses (*Adenoviridae*). Interestingly, virome

Figure 2 : Comparison of the 16 unassembled human gut viromes and the assembled dataset based on their tetranucleotide compositions. The NMDS was generated from the pairwise distances computed from the tetranucleotides frequency bias. Each virome is named from the subject (H1, H2, L1, L2) and the day of sampling (day 1, 2, 7 or 8). Samples taken from the same individual are highlighted in shades of blue, yellow and red. Highlighted in green are both the control dataset and the combined assembly.



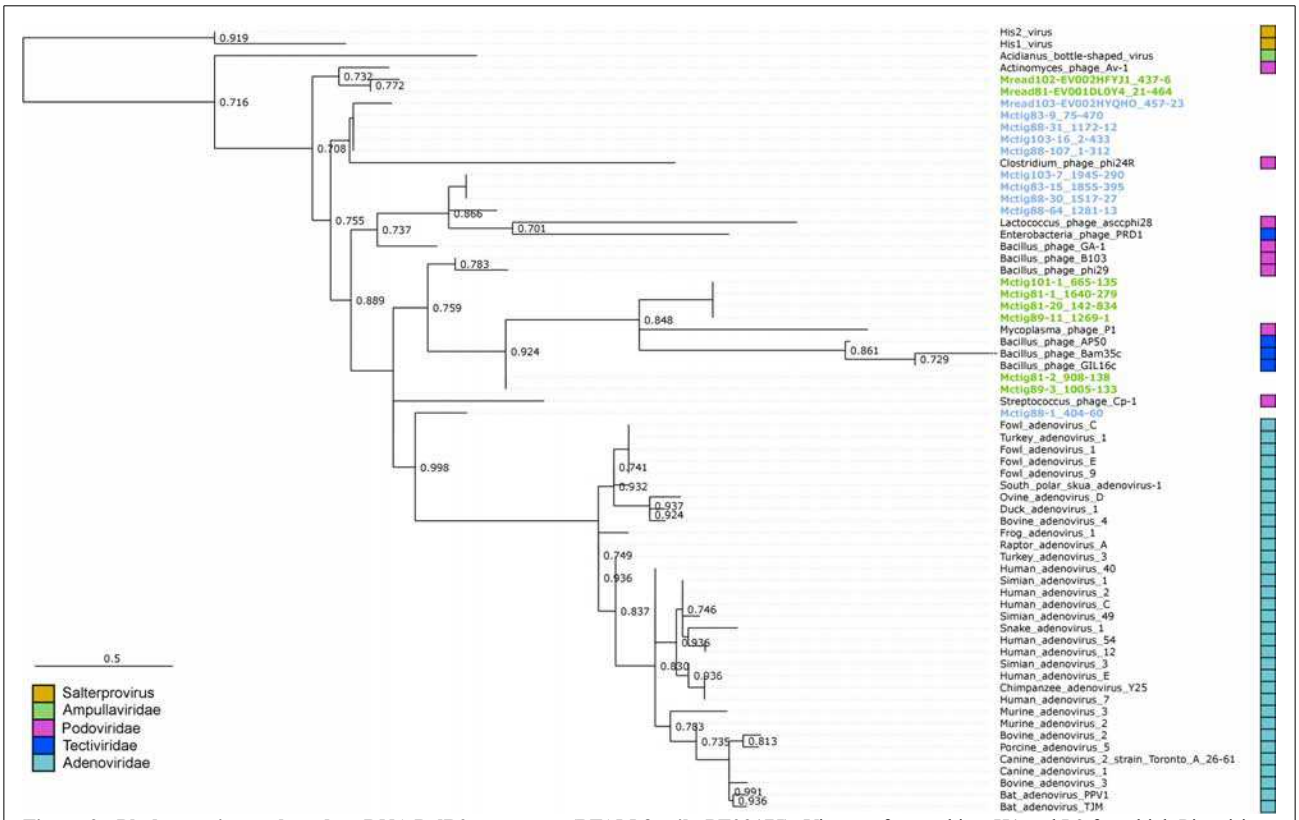


Figure 3 : Phylogenetic tree based on DNA PolB2 sequences (PFAM family PF03175). Viromes from subject H1 and L2 for which Picovirinae was the most retrieved viral family. Sequences from subject H1 and L2 are highlighted in green and blue respectively. Bootstraps scores greater than 0.70 are indicated on the tree.

sequences from each individual are gathered on the tree, highlighting that these two phage populations are distinct. Such specificity of viral strains to each individual was noted on a more general scale through virome analysis of genetically linked individuals [28]. In this example, phylogenetic analysis of an abundant viral family confirmed the conclusions drawn from the comparisons of whole viromes, highlighting the complementary aspects of these tools in better understanding such datasets.

Individual viral genome recruitment plots

In order to go beyond the analysis of single reads through BLAST or phylogenetic tools, recruitment plots of metagenomic sequences for reference genomes can be used to quantify the degree to which each viral genome is represented in a metagenome (see for example [54]). Indeed, visualizing a chosen genome and the repartition of its associated reads is useful to determine which genes of a known virus is found in an environmental dataset, and the similarity level between reference and virome sequences. Recruitment plots can be generated in Metavir, and here again, several datasets can be included in a single plot in order to compare the gene conservation of a virus between different samples. As an example, this technique was here used to further study the Lactococcus phage 1706, one of the most abundant phage in the 16 datasets from the human gut. As this phage has been isolated from bacteria involved in milk fermentation and not directly from gut microbes, its actual presence in human gut samples is questionable. The recruitment plot of Lactococcus phage 1706 shows that most characterized genes (coding for the main function of the genome, *i.e.* replication and structure module, highlighted in red on the plot) are retrieved whereas most of the unknown genes (in blue) are not (Figure 4). This indicates that even though phage 1706 is the nearest neighbor of abundant human gut phage(s) in the current state of the reference databases, said phages do not have a gene content entirely similar to phage 1706. Furthermore, a gene cassette made of two putative tail proteins and two other structural proteins known to be major players of phage-host specificity are scarcely retrieved in the different datasets ([55] ; black frame on Figure 4). Thus, it is very likely that the phages retrieved in the human gut viromes, even though similar to this *Lactococcus* phage, infect an alternative host.

Analyzing assembled datasets using the new contig section

Even though unassembled viromes proved to be useful toward a better characterization of environmental viral communities, long genomic fragments generated through the assembly of metagenomic datasets are, when available, often much more biologically informative. First, complete ORFs predicted out of such sequences are more often similar to known viruses than short reads [56], can be directly used in phylogenies providing more robust results than using reads representing only a portion of a gene, and can be used in determining the gene pool and genetic diversity of a viral community [57]. Moreover, analysis of the genomic content and architecture can provide decisive insights into virus classification and viral groups evolution [21].

A new section dedicated to the annotation and the navigation within sets of contigs has therefore been

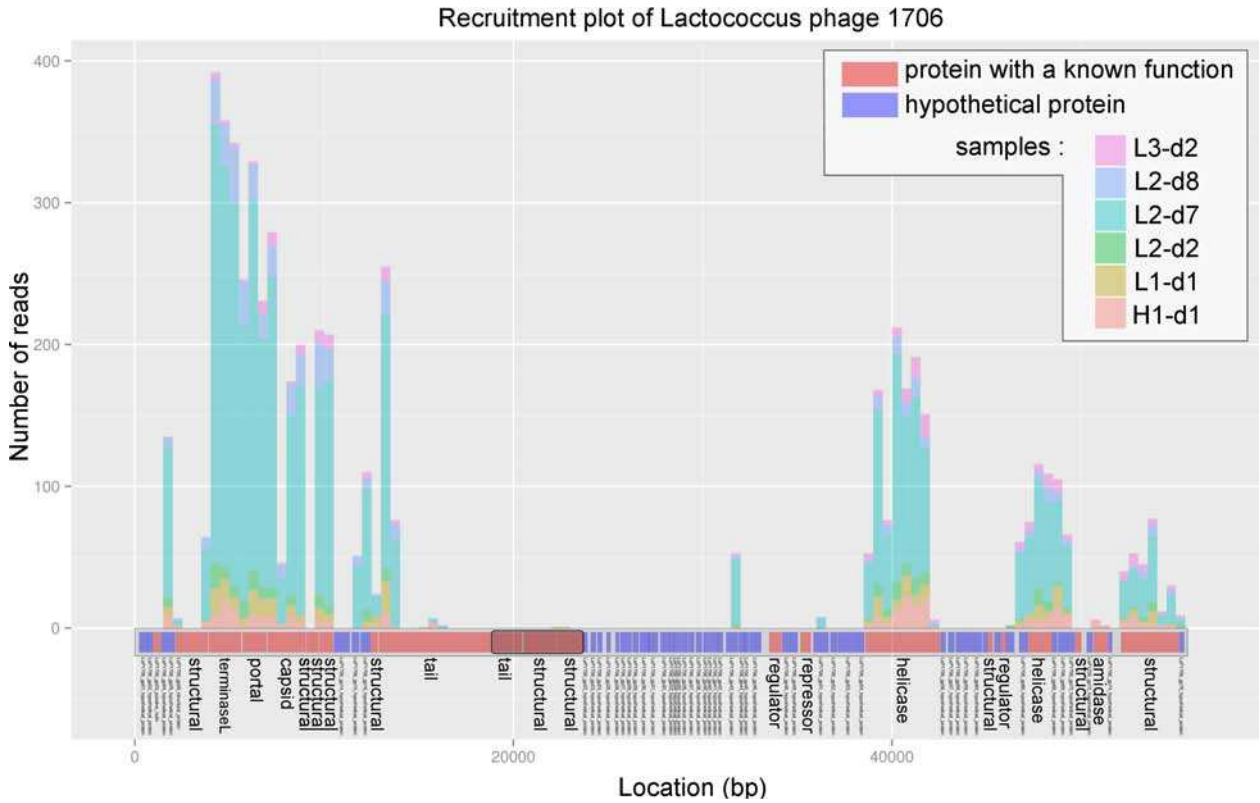


Figure 4 : Recruitment plot of Lactococcus phage 1706 for 6 human gut datasets. Only the 6 viromes with more than 200 reads recruited by this genome were included, *i.e.* reads with a best BLAST hit against this genome. Stacked histograms represent the total number of reads similar to each 500-nt long genome part, with a different color for each virome. Each gene is plotted as a rectangle on the genome map at the bottom, with hypothetical proteins in blue, and characterized genes in red. A black frame highlights the four genes involved in host specificity.

implemented in Metavir. Assembled viromes, *i.e.* sets of contigs, are first uploaded by users. Contigs are then standardly annotated : ORFs are predicted [41] and annotated using sequence similarity results against viral genomes and a protein domains. The taxonomic composition of this type of dataset is then computed using either gene or contig affiliation. In addition to this general composition, navigating through annotated contigs individually can provide valuable information. For this purpose, contig maps and annotations can be displayed for every contig. As datasets can consist of tens of thousands contigs, subsets of contigs can be selected. Users can easily choose to visualize contigs (i) longer than a defined threshold and/or (ii) predicted as circular or linear and/or (iii) affiliated to a particular viral family and/or (iv) possessing a particular gene. Finally, the different methodologies available for short read viromes are also useful for assembled datasets and were made available in Metavir. To this purpose, these tools were specifically adapted : taxonomic composition as seen above, but also phylogenies generated using predicted ORFs and genetic diversity computed using either predicted ORFs or domain conservation.

For the assembled human gut virome used as an example in this section ("Human gut - All subjects" in Metavir), 43,078 ORFs were predicted on the 10,202 uploaded contigs. Furthermore, 60 contigs were predicted as circular and represent potential complete viral genome. Using the "contig selection" panel, large contigs (>15kb) similar to Lactococcus phage 1706 were selected and further examined. For each selected contig, a summary of its annotations is available as a an interactive map. The largest sequence (contig 187_43, 60,257 bp) seems to be composed of two set of genes associated with known viral genomes (green genes at both ends of the contig), when a third and central part is made of shorter uncharacterized genes (red genes) (Figure 5). All genes but three are on the same strand (-), as generally observed in phage genomes. Moreover, no partial genes are predicted at either ends of the sequence, indicating that this contig may represent a complete genome.

Further relationships between selected contigs and viral references to which they are affiliated can then be displayed as an interactive network, where contigs and reference genomes are represented by nodes and sequence similarities as edges. For example, the network containing

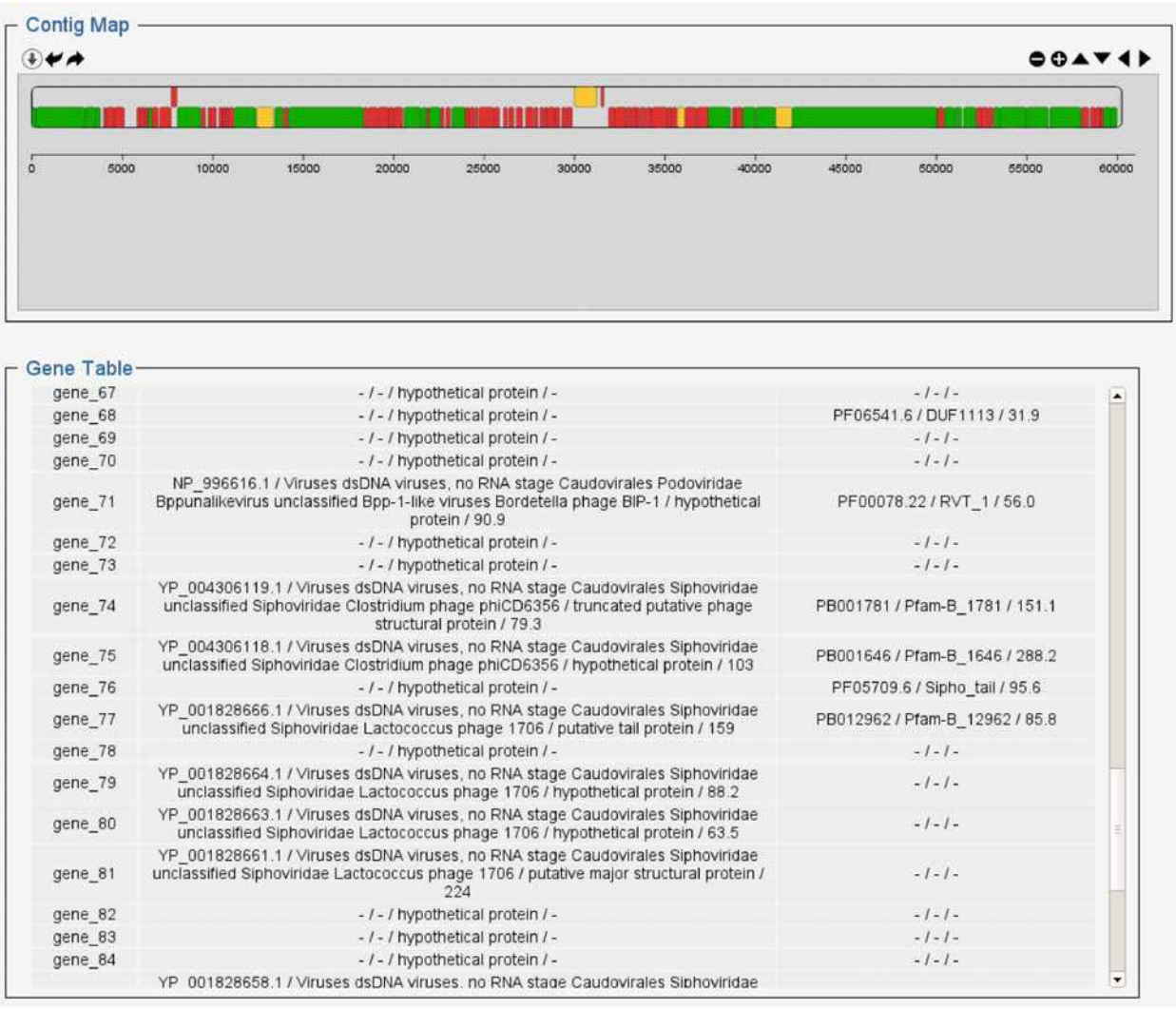


Figure 5 : Automatic annotation of contig 187_43 (complete name: 1470_2012_5M_iter1 _k63_scaffold187_43.0). On top, a dynamic map displays the predicted genes colored by affiliation (green for genes affiliated to Refseq Virus, yellow when only PFAM affiliation is available, and red for uncharacterized genes). The associated gene table (below) displays for each gene the accession number and annotation of the most similar gene Refseq and most similar domain in PFAM (when available).

contigs associated with Lactococcus phage 1706 helps to rapidly identify that these contigs are both related to each other and to several *Siphoviridae* genomes (Figure 6A). Contigs and references can then be selected in this network and a genome comparison of the chosen sequences can be displayed. This map-to-map comparison can be used to identify collinearity between different genomes or genomic fragments. When compared to the complete genome of Lactococcus phage 1706, contig 187_43 can definitely be considered as a putative complete genome closely related to this phage, as both their sizes and gene organizations are very similar. Interestingly, the similarities between this contig and the *Clostridium* phage phiCD6356 are limited to two genes which are part of the host-associated cassette previously discussed. Thus, contig 187_43 is likely originating from a phage closely related to *Lactococcus* phage 1706, but which could rather infect members of the *Clostridium* genus. The second contig displayed on figure 6B, contig 289_22.4, only shares one core gene module with phage 1706, and harbors several similarities to a distinct *Clostridium* phage. These two contigs, that both exhibit similarities to Lactococcus phage 1706, are here shown to be heterogeneous in nature. Furthermore, genes of contig 187_43 similar to Lactococcus phage 1706 correspond to the genes frequently retrieved in unassembled datasets (Figure 4), indicating that this contig might represent a prevalent virotype of the human gut. This genomic analysis of large assembled sequences exemplifies how such datasets can provide further insights into viral

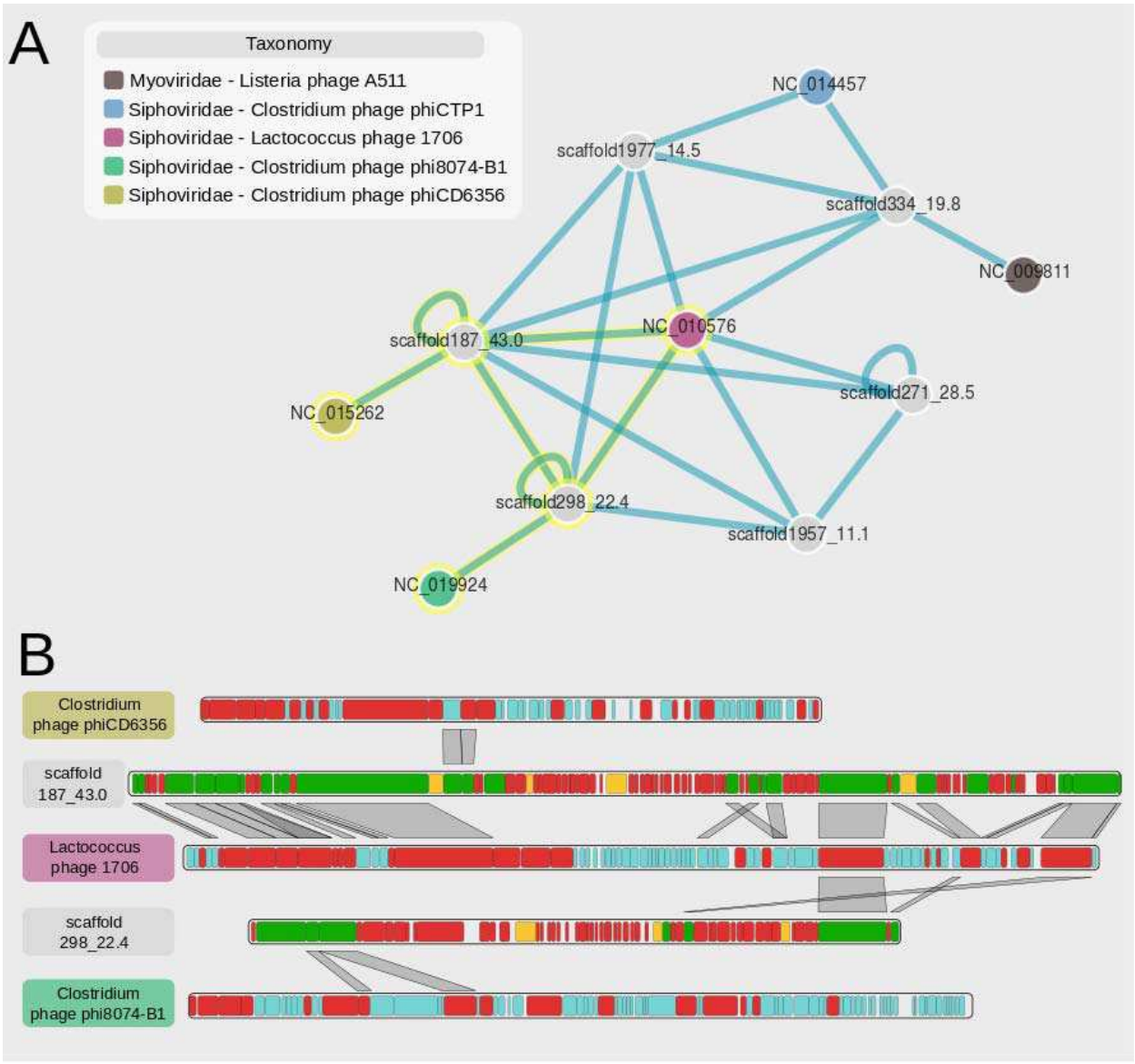


Figure 6 : Contig comparison through network and map-to-map plots. A. Contig network including 6 contigs affiliated to Lactococcus phage 1706. Each contig and reference genomes are displayed as nodes, and BLAST similarities are displayed as edges. In this network, we chose to color nodes according to the taxonomy of the reference genomes, and to keep links between nodes only when two genes or more were found to be similar between the two sequences. B. Map comparison for contigs and genomes selected on the network (highlighted in yellow in A). The maps of these five selected sequences are vertically stacked, and BLAST hits between genes of two consecutive maps are depicted with gray frames. Sequences were re-ordered to display similarities between Lactococcus phage 1706 and the two contigs, as well as similarities between these contigs and *Clostridium* phages. In both network and map comparison, the contig names were simplified: complete name of contig 187_43 is 1470_2012_5M_iter1_k63_scaffold187_43.0, contig 298_22.4 is 1470_2012_5M_iter2_k47_ scaffold298_22.4, contig 334_19.8 is 1470_2012_5M_iter2_k47_scaffold334_19.8, contig 1977_14.5 is 1470_1013_5M_iter6_k39_scaffold1977_14.5, contig 271_28.5 is 1470_2012_5M_iter2_k47_ scaffold271_28.5, and contig 1957_11.1 is 1470_1013_5M_iter6_k39_scaffold1957_11.1.

communities and viral species.

Conclusion

This new release of Metavir provides a wide range a tools to analyze either raw or assembled viral metagenomes in a comprehensive way. As virome projects now regularly encompass multiple samples and as more and more viromes are being published, a special effort was made towards virome comparison. Two new large scale methods were implemented and all existing Metavir tools were modified so that they can be used to compare datasets. Furthermore, a new section has been specifically developed to handle sets of large genomic contigs. As these datasets can be large and as all individual sequences can be of interest, we paid special attention to the interface, with filtering panels and network visualization. Selected contigs can then be analyzed in detail by comparing their automatic annotations in terms of gene content and genomic maps. Finally, with its extended or new tools and sections, Metavir 2 provides a comprehensive framework with a user-friendly interface to explore any kind of viromes, and should help virologists to make the most of their metagenomics data.

Availability and requirements

Project Name: Metavir
Project home page: http://metavir-meb.univ-bpclermont.fr
Operating system(s): Linux, Mac Os X, Microsoft Windows;
Programming language: Perl, Php, Javascript, Css, R;
Licence: GPL3;
Any restrictions to use by non-academics: No.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SR, FE and DD designed the tools. SR, JT and AM developed the different scripts. SR and FE wrote the manuscript.

Acknowledgements

SR was supported by a PhD grant from the French defense procurement agency (DGA, Direction Générale de l'Armement). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript .

References

1. Suttle CA: **Viruses in the sea.** *Nature* 2005, **437**:356–61.

2. Delwart EL: **Viral metagenomics**. *Reviews in Medical Virology* 2007, **17**:115–131.

3. Fancello L, Raoult D, Desnues C: **Computational tools for viral metagenomics and their application in clinical research.** *Virology* 2012, **434**:162–174.

4. Suttle CA: **Marine viruses--major players in the global ecosystem.** *Nature Reviews Microbiology* 2007, **5**:801–812.

5. Rohwer F, Thurber RV: **Viruses manipulate the marine environment.** *Nature* 2009, **459**:207–212.

6. Hatfull GF, Hendrix RW: **Bacteriophages and their Genomes**. *Current Opinion in Virology* 2011, **1**:298–303.

7. Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D: **Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics**. *PLoS One* 2012, **7**:e33641.

8. Duhaime MB, Sullivan MB: **Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline.** *Virology* 2012, **434**:181–186.

9. Vega Thurber R, Haynes M, Breitbart M, Wegley L, Rohwer F: **Laboratory procedures to generate viral metagenomes.** *Nature protocols* 2009, **4**:470–483.

10. Willner D, Hugenholtz P: **From deep sequencing to viral tagging: Recent advances in viral metagenomics.** *BioEssays : news and reviews in molecular, cellular and developmental biology* 2013:1–7.

11. Fancello L, Trape S, Robert C, Boyer M, Popgeorgiev N, Raoult D, Desnues C: **Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara.** *The ISME journal* 2012:1–11.

12. Whon TW, Kim M-S, Roh SW, Shin N-R, Lee H-W, Bae J-W: **Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere**. *Journal of virology* 2012, **86**:8221–8331.

13. Kristensen DM, Mushegian AR, Dolja V V, Koonin E V: **New dimensions of the virus world discovered through metagenomics.** *Trends in microbiology* 2010, **18**:11–19.

14. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan P, Hui J, Marshall J, Simons JF, Egholm M, Paddock CD, Shieh W, Goldsmith CS, Zaki SR, Catton M, Lipkin WI: **A new arenavirus in a cluster of fatal transplant-associated diseases**. *The New England Journal of Medicine* 2008, **358**:991–998.

15. Koren S, Treangen TJ, Pop M: **Bambus 2: scaffolding metagenomes.** *Bioinformatics* 2011, **27**:2964–71.

16. Peng Y, Leung HCM, Yiu SM, Chin FYL: **IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth**. *Bioinformatics* 2012, **28**:1420–1428.

17. Minot S, Wu GD, Lewis JD, Bushman FD: **Conservation of Gene Cassettes among Diverse Viruses of the Human Gut**. *PLoS ONE* 2012, **7**:e42342.

18. Ng TFF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C, Anthony S, Ruan Y, Rohwer F, Breitbart M: **Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes.** *PLoS One* 2011, **6**:e20579.

19. Rosario K, Duffy S, Breitbart M: **Diverse circovirus-like genome architectures revealed by environmental metagenomics.** *Journal of general virology* 2009, **90**:2418–2424.

20. Diemer GS, Stedman KM: **A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses.** *Biology direct* 2012, **7**:13.

21. Roux S, Krupovic M, Poulet A, Debroas D, Enault F: **Evolution and diversity of the Microviridae viral family through a collection of 81 new complete genomes assembled from virome reads.** *PloS one* 2012, **7**:e40418.

22. Coetzee B, Freeborough M-J, Maree HJ, Celton J-M, Rees DJG, Burger JT: **Deep sequencing analysis of viruses infecting grapevines: Virome of a vineyard.** *Virology* 2010, **400**:157–63.

23. Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF: **Metagenomic assembly reveals dynamic viral populations in hypersaline systems.** *Applied and environmental microbiology* 2012, **78**:6309 – 6320.

24. Minot S, Grunberg S, Wu GD, Lewis JD, Bushman FD: **Hypervariable loci in the human gut virome.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109**:3962–6.

25. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F: **Metavir: a web server dedicated to virome analysis.** *Bioinformatics* 2011, **27**:3074–3075.

26. Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S, Furman M, Jamindar S, Nasko DJ: **VIROME: a standard operating procedure for analysis of viral metagenome sequences**. *Standards in Genomic Sciences* 2012, **6**:427–439.

27. Lorenzi H a, Hoover J, Inman J, Safford T, Murphy S, Kagan L, Williamson SJ: **TheViral MetaGenome Annotation Pipeline(VMGAP):an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data.** *Standards in genomic sciences* 2011, **4**:418–29.

28. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, Rohwer F, Gordon JI: **Viruses in the faecal microbiota of monozygotic twins and their mothers.** *Nature* 2010, **466**:334–338.

29. Ray J, Dondrup M, Modha S, Steen IH, Sandaa R-A, Clokie M: **Finding a needle in the virus metagenome haystack--micro-metagenome analysis captures a snapshot of the diversity of a**

bacteriophage armoire. *PloS one* 2012, **7**:e34238.

30. Li R, Li Y, Kristiansen K, Wang J: **SOAP: short oligonucleotide alignment program.** *Bioinformatics* 2008, **24**:713–4.

31. Namiki T, Hachiya T, Tanaka H, Sakakibara Y: **MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads.** *Nucleic acids research* 2012, **40**:e155.

32. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R, Antonopoulos DA, Barott K, Cottrell MT, Desnues C, Dinsdale EA, Furlan M, Haynes M, Henn MR, Hu Y, Kirchman DL, McDole T, McPherson JD, Meyer F, Miller RM, Mundt E, Naviaux RK, Rodriguez-Mueller B, Stevens R, Wegley L, Zhang L, Zhu B, Rohwer F: **The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes.** *PLoS computational biology* 2009, **5**:e1000593.

33. Ondov BD, Bergman NH, Phillippy AM: **Interactive metagenomic visualization in a Web browser.** *BMC bioinformatics* 2011, **12**:385.

34. Willner D, Thurber RV, Rohwer F: **Metagenomic signatures of 86 microbial and viral metagenomes.** *Environmental microbiology* 2009, **11**:1752–1756.

35. R Core Team: [http://www.R-project.org] *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013.

36. Suzuki R, Shimodaira H: **Pvclust: an R package for assessing the uncertainty in hierarchical clustering.** *Bioinformatics* 2006, **22**:1540–1542.

37. Oksanen J, Kindt R, Legendre P, O'Hara B, Simpson GL, Solymos P, Stevens MHH, Wagner H: *The vegan Package*. 2008.

38. Price MN, Dehal PS, Arkin AP: **FastTree 2--approximately maximum-likelihood trees for large alignments**. *PLoS One* 2010, **5**:e9490.

39. Smits S a, Ouverney CC: **jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web.** *PloS one* 2010, **5**:e12267.

40. Wickham H: *ggplot2: Elegant Graphics for Data Analysis*. Springer Publishing Company; 2009.

41. Noguchi H, Taniguchi T, Itoh T: **MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes.** *DNA research* 2008, **15**:387–96.

42. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215**:403–410.

43. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD: **The Pfam protein families database.** *Nucleic acids research* 2012, **40**:D290–301.

44. Eddy SR: **Accelerated Profile HMM Searches.** *PLoS computational biology* 2011, **7**:e1002195.

45. Edgar RC: **Search and clustering orders of magnitude faster than BLAST.** *Bioinformatics (Oxford, England)* 2010, **26**:2460–1.

46. Lopes CT, Franz M, Kazi F, Donaldson SL, Morris Q, Bader GD: **Cytoscape Web: an interactive web-based network browser.** *Bioinformatics* 2010, **26**:2347–8.

47. Rutherford K, Parkhill J, Crook J, Horsnell T, Barrell B, Rice P: **Artemis: sequence visualization and annotation**. *Bioinformatics* 2000, **16**:944–945.

48. Sullivan MJ, Petty NK, Beatson S a: **Easyfig: a genome comparison visualizer.** *Bioinformatics* 2011, **27**:1009–10.

49. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, Lewis JD, Bushman FD: **The human gut virome: inter-individual variation and dynamic response to diet**. *Genome Research* 2011, **21**:1616–1625.

50. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, Chan AM, Haynes M, Kelley S, Liu H, Mahaffy JM, Mueller JE, Nulton J, Olson R, Parsons R, Rayhawk S, Suttle CA, Rohwer F: **The marine viromes of four oceanic regions.** *PLoS biology* 2006, **4**:e368.

51. Rosario K, Nilsson C, Lim YW, Ruan Y, Breitbart M: **Metagenomic analysis of viruses in reclaimed water.** *Environmental microbiology* 2009, **11**:2806–20.

52. Yoshida M, Takaki Y, Eitoku M, Nunoura T, Takai K: **Metagenomic Analysis of Viral Communities in (Hado) Pelagic Sediments**. *PLOS ONE* 2013, **8**:e57271.

53. Edwards RA, Rohwer F: **Viral metagenomics.** *Nature Reviews Microbiology* 2005, **3**:504–510.

54. Ghai R, Martin-Cuadrado A-B, Molto AG, Heredia IG, Cabrera R, Martin J, Verdú M, Deschamps P, Moreira D, López-García P, Mira A, Rodriguez-Valera F: **Metagenome of the Mediterranean deep chlorophyll maximum studied by direct and fosmid library 454 pyrosequencing.** *The ISME journal* 2010, **4**:1154–66.

55. Garneau JE, Tremblay DM, Moineau S: **Characterization of 1706, a virulent phage from Lactococcus lactis with similarities to prophages from other Firmicutes.** *Virology* 2008, **373**:298–309.

56. Wommack KE, Bhavsar J, Ravel J: **Metagenomics: read length matters.** *Applied and environmental microbiology* 2008, **74**:1453–1463.

57. Hurwitz BL, Sullivan MB: **The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology**. *PLoS One* 2013, **8**:e57355.

Bilan et perspectives pour le serveur Metavir

Utilisation et valorisation de l'outil

Même si les études de métagénomique virale sont encore relativement peu nombreuses, notamment en regard des études de diversité des communautés de micro-organismes dans l'environnement, le serveur Metavir a enregistré une croissance régulière du nombre de projets déposés et du nombre de paires de bases étudié (Figure I.2). La section dédiée aux séquences assemblées, mise en ligne à l'automne 2012, a rapidement été sollicitée par les différents utilisateurs, ce qui explique l'augmentation plus limitée de la quantité de séquence étudiées par rapport au nombre de projets. Ainsi, on retrouve par l'intermédiaire des projets déposés sur Metavir cette tendance des études de viromes vers l'étude de séquences assemblées, pour lesquels il y a moins de données brutes (moins de paires de bases), mais au final plus d'information biologique.

L'article décrivant le serveur Metavir a tout d'abord été cité au sein d'articles décrivant les protocoles d'obtention des viromes (Duhaime & Sullivan, 2012; Solonenko *et al.*, 2013), ou les processus d'analyse de métagénomes viraux (Fancello *et al.*, 2012a; Reyes *et al.*, 2012; Wylie *et al.*, 2012), avant que de véritables analyses de viromes utilisant le serveur ne soient effectivement publiées (McDaniel *et al.*, 2013; Pérez-Brocal *et al.*, 2013; Yoshida *et al.*, 2013).

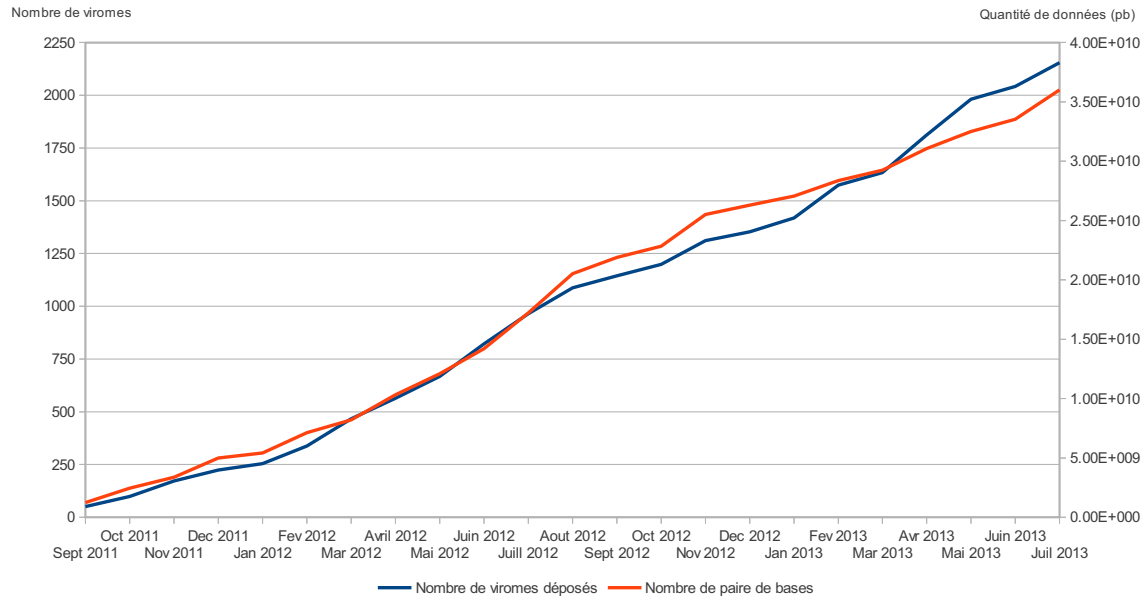


Figure I.2 : Nombre de projets déposés sur le serveur Metavir et nombre de paires de bases associées. *Les échelles en ordonnées correspondent à gauche au nombre cumulé de viromes, et à droite à la quantité totale de nucléotides analysés.*

Limites et futurs développements

Plusieurs sections de Metavir pourraient encore être améliorées et adaptées aux nouveaux types de données. Tout d'abord, l'organisation et la navigation au sein des différents viromes, notamment l'ensemble des projets publics, est de plus en plus difficile de par l'augmentation du nombre de projets disponibles. Si la structure mise en place convenait au lancement du serveur, il sera certainement nécessaire de modifier cette dernière afin de mieux présenter et identifier les jeux de données disponibles. Les métadonnées associées à ces projets sont également trop sporadiques en l'état actuel du serveur. L'utilisation d'ontologie (comme par exemple l' "*Environmental Ontology*") pourrait faciliter à la fois l'organisation des projets et l'inscription comme la consultation de ces métadonnées.

L'utilisation de la seule base de donnée RefseqVirus, réduite uniquement aux génomes viraux complets, limite l'annotation potentielle des séquences de viromes, mais est contrainte par des problématiques de capacité de calcul. En effet, pour être le plus exhaustif possible, il est souvent procédé lors des études métagénomiques à un BLAST contre une base de données généraliste (de type NR, disponible au NCBI) mais ce type de comparaison dépasse les capacités disponibles actuellement pour le serveur Metavir en terme de temps de calcul. En effet, l'augmentation de la taille de ces bases de données généralistes, composées à présent de plusieurs centaines de millions de séquences, rend nécessaire l'utilisation de capacités de calcul exceptionnelles de type cluster de calcul géant ou grille de calcul dès lors qu'il s'agit de traiter un ensemble conséquent de jeux de données.

Il est également à noter que l'ensemble de gènes marqueurs proposé reste encore trop hétérogène, avec certains groupes pour lesquels plusieurs marqueurs sont disponibles, et d'autres pour lesquels la procédure phylogénétique n'est pas disponible. Une refonte de ces marqueurs, basée sur une analyse des génomes complets disponibles et une procédure normalisée pour la définition des alignements de référence permettrait sans doute de mieux exploiter cet outil.

Enfin, si les courbes de raréfaction et les arbres phylogénétiques permettent de dégager des tendances et de confirmer ou infirmer des hypothèses, l'utilisation d'indices statistiques permettrait de les valider plus rigoureusement. Pour les arbres phylogénétiques, il pourrait s'agir de distances et mesures de type Unifrac lors de la comparaison d'environnement, tandis que des indices de richesse comme l'entropie de Shannon et l'indice Schao pourraient être appliqués aux résultats de clusterisation.

Le développement de Metavir s'est ainsi inscrit dans un processus plus large de création d'outils boinformatiques pour les métagénomes et pour les viromes. Concernant la

métagénomique virale, une procédure normalisée d'annotation des séquences a été décrite (Lorenzi *et al.*, 2011), et un autre serveur d'analyse a été mis en ligne à l'été 2012 : Virome.org (Wommack *et al.*, 2012). Ce serveur propose un traitement des séquences en trois grandes parties : un nettoyage des séquences, une comparaison à un ensemble exhaustif de bases de données, et enfin une caractérisation de chaque séquence du virome. Cette dernière étape constitue la spécificité principale de ce serveur d'analyse, puisqu'elle propose pour chaque séquence une synthèse des comparaisons réalisées à la fois contre les bases de données de référence et contre une base de donnée métagénomiques (MGOL, pour MetaGenomes On-Line). Un arbre de décision a ainsi été réalisé, permettant d'aboutir à une affiliation automatique la plus précise possible. Une interface complète a de plus été développée pour permettre la sélection, consultation et extraction des résultats de ces affiliations. Toutefois, les seules comparaisons de viromes proposées par ce serveur sont basées sur les différentes affiliations taxonomiques et fonctionnelles. De même, Virome.org ne propose pas d'outil pour caractériser la richesse génétique d'un jeu de données, pour compléter et dépasser les affiliations de séquences individuelles par meilleur résultat de BLAST, ou encore pour analyser des séquences assemblées. Ainsi, Virome.org consiste avant tout en une série de comparaisons à différentes bases de données, mais ne permet pas de mener l'analyse globale d'un métagénome viral.

Metavir constitue donc actuellement un ensemble unique et cohérent d'outils bioinformatiques complémentaires dédiés à l'analyse de viromes. Ces outils sont adaptés aux variations tant qualitatives que quantitatives entre les jeux de données issus des différentes techniques de séquençage à haut-débit, et peuvent être utilisés quel que soit l'objectif de l'étude entreprise.

Chapitre II – Potentiel fonctionnel des génomes viraux

L'une des conséquences du faible nombre de génomes viraux actuellement séquencés est une méconnaissance du potentiel fonctionnel global des communautés virales. Or, cette question touche à l'essence même des virus et à leur définition. D'abord considérés comme agents chimiques capable de reproduction, le fait que du matériel génétique soit bien le support de cette reproduction les a fait basculer vers le monde du vivant (Van Regenmortel, 2003). Toutefois, les virus sont toujours largement considérés comme des entités inertes du point de vue métabolique, plus proches de parasites génétiques que de véritables entités douées de vie (Lopez-Garcia, 2012).

Cette vision des virus a cependant été remise en question par la découverte au sein de génomes de phages de gènes codant pour des protéines impliqués dans des métabolismes cellulaires tels que la photosynthèse, dont la seule présence est ainsi apparue comme fondamentalement contradictoire avec la nature inerte des virus (Lindell *et al.*, 2004). Ces observations ont été complétées par la mise en lumière de différents cas de transferts de gènes métaboliques entre les génomes de virus et de leurs hôtes, au niveau des phages bactériens (Rohwer *et al.*, 2000; Sullivan *et al.*, 2005) mais aussi dans le cas de virus d'eucaryotes (Monier *et al.*, 2009; Moreau *et al.*, 2010). En outre, il a été montré (notamment à partir de l'étude de cyanophages) que ces gènes de métabolisme identifiés au sein des génomes viraux sont effectivement actifs et transcrits au cours du cycle viral, et influencent ainsi le métabolisme de la cellule hôte (Lindell *et al.*, 2005, 2007). Il reste cependant à déterminer s'il s'agit de cas isolés, ou si la présence de gènes de métabolismes potentiellement actifs durant le cycle infectieux est effectivement courante dans l'ensemble du monde viral.

Méta-analyse fonctionnelle de viromes

Dans ce contexte, les approches métagénomiques constituent une possibilité intéressante d'étude du potentiel fonctionnel de communautés complexes. Ainsi, en 2008, une méta-analyse de 45 métagénomes microbiens et 42 viromes a été publiée, comparant les métagénomes entre eux en fonction du type d'écosystème étudié (Dinsdale *et al.*, 2008). Cette étude a notamment permis de montrer que les communautés virales et microbiennes d'écosystèmes proches possédaient un ensemble de gènes fonctionnels relativement similaires (Figure In.8). De plus, un résultat plus inattendu a pu être observé : le potentiel fonctionnel identifié au sein des viromes était quasiment identique à celui observé au sein des métagénomes microbiens. Ce résultat a par la suite été repris par d'autres auteurs, notamment Kristensen et collaborateurs qui ont souligné l'éventuel biais de cette étude lié à la

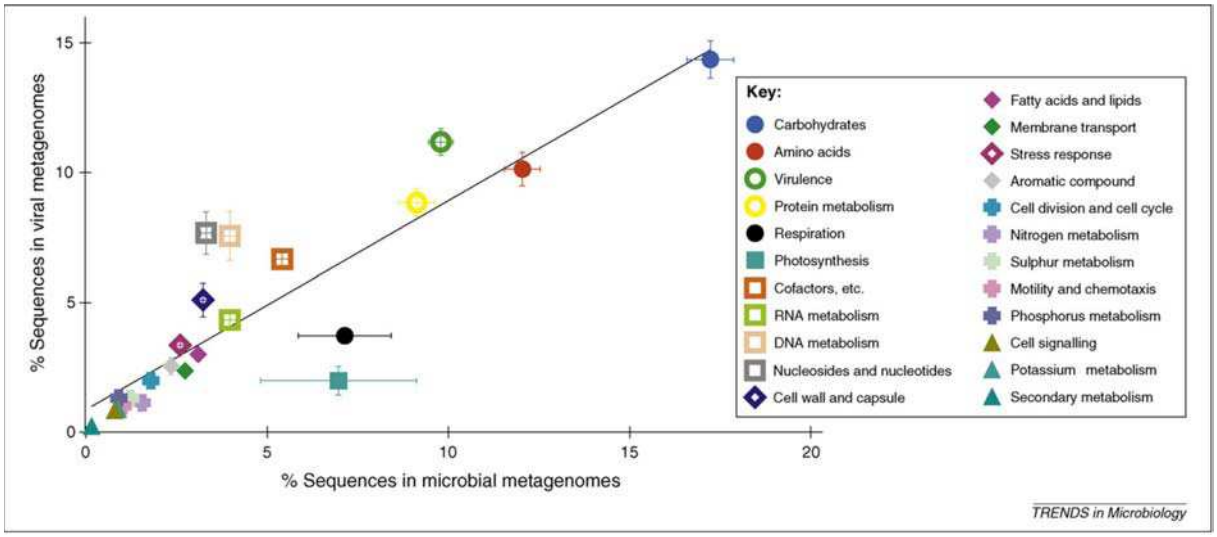


Figure II.1 : Comparaison des affiliations fonctionnelles de 45 microbiomes et 42 viromes (respectivement en abscisse et ordonnée, figure issue de Kristensen* et al.*, 2010). *Une moyenne du ratio d'affiliation à chaque catégorie fonctionnelle est indiquée pour chaque type de métagénome, ainsi que les valeurs maximales et minimales.*

contamination potentielle de viromes par du matériel génétique cellulaire (Figure II.1 ; (Kristensen *et al.*, 2010)).

De manière plus générale, ces résultats mettent en avant le problème d'interprétation des résultats d'analyse de viromes, et notamment le nombre important de séquences détectées comme similaires à des gènes microbiens. Deux hypothèses peuvent expliquer ces similarités entre séquences de viromes et génomes cellulaires :

- Il s'agit de véritables séquences d'origine virale, pouvant être issues d'échanges ponctuels de gènes entre génomes microbiens et viraux, et/ou de l'intégration de génomes viraux sous forme de prophages au sein du génome de l'hôte.
- Il s'agit d'une contamination du virome par des séquences d'origine cellulaire, soit liée à une contamination *stricto sensu* de l'échantillon, soit associée à la présence de GTA (Agents de Transfert de Gène), structures similaires aux capsides virales mais comprenant de l'ADN microbien.

Une méta-analyse de 67 viromes a ainsi été menée afin de déterminer l'origine de ces similarités entre séquences de viromes et génomes cellulaire, et *in fine* d'estimer le plus justement et rigoureusement possible le véritable potentiel fonctionnel des génomes viraux.

Article III

Uncontaminated viromes reveal the abundance and diversity of metabolism genes in environmental viruses.

Roux Simon[1,2], Krupovic Mart[3], Debroas Didier[1,2], Forterre Patrick[3,4], Enault François[1,2]

[1] Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", Clermont-Ferrand , France

[2] CNRS UMR 6023, LMGE, Aubière, France

[3] Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, Paris, France

[4] Laboratoire de Biologie Moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Orsay, France

Soumis à **The ISME Journal**

Matériel supplémentaire : Annexe A.4

Uncontaminated viromes reveal the abundance and diversity of metabolism genes in environmental viruses.

Roux Simon[1,2], Krupovic Mart[3], Debroas Didier[1,2], Forterre Patrick[3,4], Enault François* [1,2]

[1] Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", Clermont-Ferrand , France [2] CNRS UMR 6023, LMGE, Aubière, France [3] Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, Paris, France [4] Laboratoire de Biologie Moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Orsay, France

Keywords : virus, functional metagenomics

Abstract

Although the importance of viruses in natural ecosystems is widely acknowledged, the functional potential of viral communities is yet to be determined. Viral genomes are traditionally believed to carry only those genes that are directly pertinent to the viral life cycle. This general view has been recently challenged by the discovery of metabolism genes, such as photosystem or phosphate metabolism genes, in cyanophages. Metagenomic approaches made it possible to extend these results to a community scale, and several studies concluded that microbial and viral communities encompass similar functional potentials. However, these conclusions could originate from the presence of cellular DNA within viral metagenomes. To identify and characterize this putative cellular genetic material in viromes, we first searched for the presence of ribosomal gene sequences in a set of 67 published viromes sampled from all types of biomes, and developed a computational method to estimate the proportion and origin of cellular sequences in each virome. In a few cases, Gene Transfer Agents (GTA, host-encoded virus-like particles that package DNA fragments of the host chromosome) were found to be directly responsible for the observed impurities. When considering only viromes for which no cellular DNA was detected, the functional potential of viral and microbial communities was found to be fundamentally different. Yet, viruses still appear to harbor a substantial number of genes associated with carbon and energy metabolism as well as genes for amino-acid metabolism and transport, and, unexpectedly, ribosomal proteins. Thus, even though previous conclusions on the equal functional potential of viruses and cellular organisms can be dismissed, the presence of auxiliary genes involved in various metabolic pathways within viral genomes seems to be a general trend in the virosphere.

Introduction

Studies of the quantitative and functional importance of viruses in natural environments have emerged more than 20 years ago with the report on high concentration of bacteriophages in natural waters (Bergh *et al.* 1989). Viruses were progressively shown to be the most abundant biological entities in the biosphere (Suttle 2007) and these observations have prompted scientists to determine the roles of viruses in diverse ecosystems. Viruses are now considered as an important factor in the control of microorganisms in various ecological niches (Wommack & Colwell 2000; Rodriguez-Valera *et al.* 2009), interfering with major biogeochemical cycles (Suttle 2007). In addition, viruses also mediate genetic exchange among bacteria by transduction (*i.e.* the process by which DNA is transferred from one bacterium to another by a virus) and have a great influence on the evolution of cellular organisms since the beginning of the cellular life (Forterre 2006). However, the extent of the functional potential encompassed in environmental viral communities is still scantily characterized.

Although viruses were first believed to carry only those genes that are directly involved in viral reproduction (Dimijian 2000), accumulation of complete viral genome sequences during the past decade revealed a deviation from this general paradigm. Besides the bona fide viral genes (*i.e.*, for virion structure and assembly, and genome replication), several viruses were found to contain "auxiliary metabolism genes". Phosphate metabolism-associated genes were, for example, described in *Roseobacter* phage SIOI (Rohwer *et al.* 2000), while several photosystem genes were discovered in cyanophages (Lindell *et al.* 2004). More recently, a eukaryotic virus was found to encode a complete sphingolipid production pathway (Monier *et al.* 2009). The discovery of such metabolism genes in viral genomes was one of the elements fueling the recently renewed debate about the true nature of viruses and their place among cellular life forms (López-García & Moreira 2009; Ludmir & Enquist 2009).

Overall, the growing awareness of the central role played by viruses in different ecosystems has ignited an interest of many scientists towards the structure and dynamics of viral communities from various biomes. As a result, a great deal of metagenomic data is currently available on diverse uncultivated viral communities. These numerous viral metagenomes (*i.e.* viromes) should provide valuable information on functional viral genes and on genes shared by viruses and microbes at a community scale.

A puzzling and recurrent observation in analyses of virome data is that the majority of reads have no similarity to any known viral sequence (Edwards &

Rohwer 2005). Several factors were proposed as plausible explanations for this low ratio of viral genome-viral metagenome matches, including the great heterogeneity in genome content and organization, the huge genetic diversity of viruses, the paucity of data on viral genomes (Allen & Wilson 2008), the absence of close references, as well as the short length of virome reads (Wommack *et al.* 2008). The direct impact of two of these factors was recently confirmed: newly sequenced viral genomes (particularly from natural ecosystems) and longer reads obtained with new pyrosequencing technologies (read length > 400 bp) have allowed assignment of more reads to viruses (Minot *et al.* 2011; Roux *et al.* 2012).

Even more surprisingly, virome reads with detectable homologues are mostly affiliated to prokaryotic genes (Edwards & Rohwer 2005; Ng *et al.* 2011; Roux *et al.* 2012) and this observation is thought to be due to (i) protein conservation across viral and cellular genomes (ii) availability of more microbial than viral sequence data, and (iii) prophage sequences within microbial genomes (Vega Thurber *et al.* 2008). Even though such reasoning explains why some reads of viral origin are affiliated to bacteria, it could not be sufficient to explain the high proportion of reads similar to bacterial genomes for some viromes.

However, these viral metagenomic data were obtained after particle isolation based on density or/and size. The viromes can therefore contain a possible non-negligible presence of cellular DNA, potentially originating from GTA (Kristensen *et al.* 2010) or DNA-containing membrane vesicles (Forterre *et al.* 2013) that might not be removed and identified by the current methods. These assumptions are consistent with the following observations: (i) only a small fraction of virome reads are significantly similar to known viruses, (ii) the detectable homologues of virome reads are primarily bacterial genes (Edwards & Rohwer 2005) and, (iii) these viral homologues represent all bacterial functions (Dinsdale *et al.* 2008).

In this study, 67 published viromes from various biomes (Table S1) were analyzed to identify and quantify the extent and possible origins of bacterial-like sequences. More than half of the analyzed datasets (43) were found to contain various amounts of cellular sequences (including GTA), and 9 correspond to viromes *sensu stricto* (*i.e.* sequence datasets exclusively from the viral community). This analysis allowed us to uncover a more accurate picture of the prevalence of diverse metabolism genes encoded by viruses, providing a first unbiased view of the functional potential of viral communities across various biomes.

Results & Discussion

What is the extent of microbial DNA presence in viromes ?

The presence of typical prokaryotic genes in viromes is an indication that the dataset might contain DNA of cellular origin. Genes coding for the ribosomal RNA (rDNA) represent a particularly good marker of such presence as no such gene was ever detected in a viral genome. The ratio of rDNA genes detected, roughly reflecting the relative proportion of cellular sequences within a given virome, ranged from 0 up to 5.3 ‰ (for an average of 2.2 ‰ for microbial metagenomes, Table S2). According to this ratio, viromes were separated into three groups (Fig. 1A, Table S2):

- Viromes with no rDNA genes detected that can be considered as devoid of cellular sequences (depicted as green in Figure 1). This dataset is composed of 9 viromes, all sampled in aquatic environments (2 from freshwater, 1 from microbialites, 3 from seawater, and 3 from hypersaline systems). Notably, all of the 21 published human gut viromes were found to contain rDNA sequences.
- Viromes with a rDNA ratio lower than 0.2‰ (2 from 10, 000 sequences), for which the amount of cellular sequences can be considered as very low and likely to be negligible (depicted in orange in Fig.1).
- Viromes with a rDNA ratio higher than 0.2‰ (up to 5.3‰) with a non negligible proportion of cellular sequences (depicted in red in Fig.1).

It has to be noted that the threshold of 0.2‰ used here to separate viromes is arbitrary, and is used to distinguish between the viromes that contain only a few sequences of cellular origin, and those that contain a fair amount of such sequences.

The detection of rDNA genes can be complemented by a more general Microbial Hit Ratio, the ratio of virome reads with a hit against a microbial genome (hereafter named MHR). Considering virome reads truncated to 100 bp (to avoid bias due to differences in read length among viromes), MHR exhibited a great variability, ranging from 0.24 to 40.31 % (6.2 % on average, Table S2). Furthermore, viromes with a “high” rDNA ratio (>0.2‰) exhibited a higher MHR (ratio of virome reads with a significant similarity to at least one prokaryotic genome) than viromes with rDNA ratio lower than 0.2‰. These two independent yet correlated results (high rDNA ratio associated with high MHR, and vice versa) points toward the presence of DNA from cellular origin within these viromes.

Anyhow, viromes with low MHR still exhibit more similarity to cellular than to viral genomes and this affiliation paradox prompted us to further investigated these cellular-like sequences. As genomic studies revealed that prophages are highly prevalent in many and diverse prokaryotes (Casjens 2003), we hypothesize that a large share of metagenomic reads from viral origin that are similar to bacterial genomes are in fact similar to prophages.

To verify this assumption, we first identified prophage-like regions in prokaryotic genomes. An in-house developed program allowed us to detect 55,837 prophage-like regions in the 1312 genomes analyzed (42 regions per genome, on average) which encompassed 11 % of the genes in the considered genomes. Once identified, virome reads similar to the prophage-like regions were determined and a Prophage Hit Ratio (PHR) was calculated as the number of hits detected within prophage-identified region divided by the total number of hits to cellular genomes for the virome. Thus, when the MHR can be seen as addressing the question "how many sequences of a virome match a cellular genome", the PHR

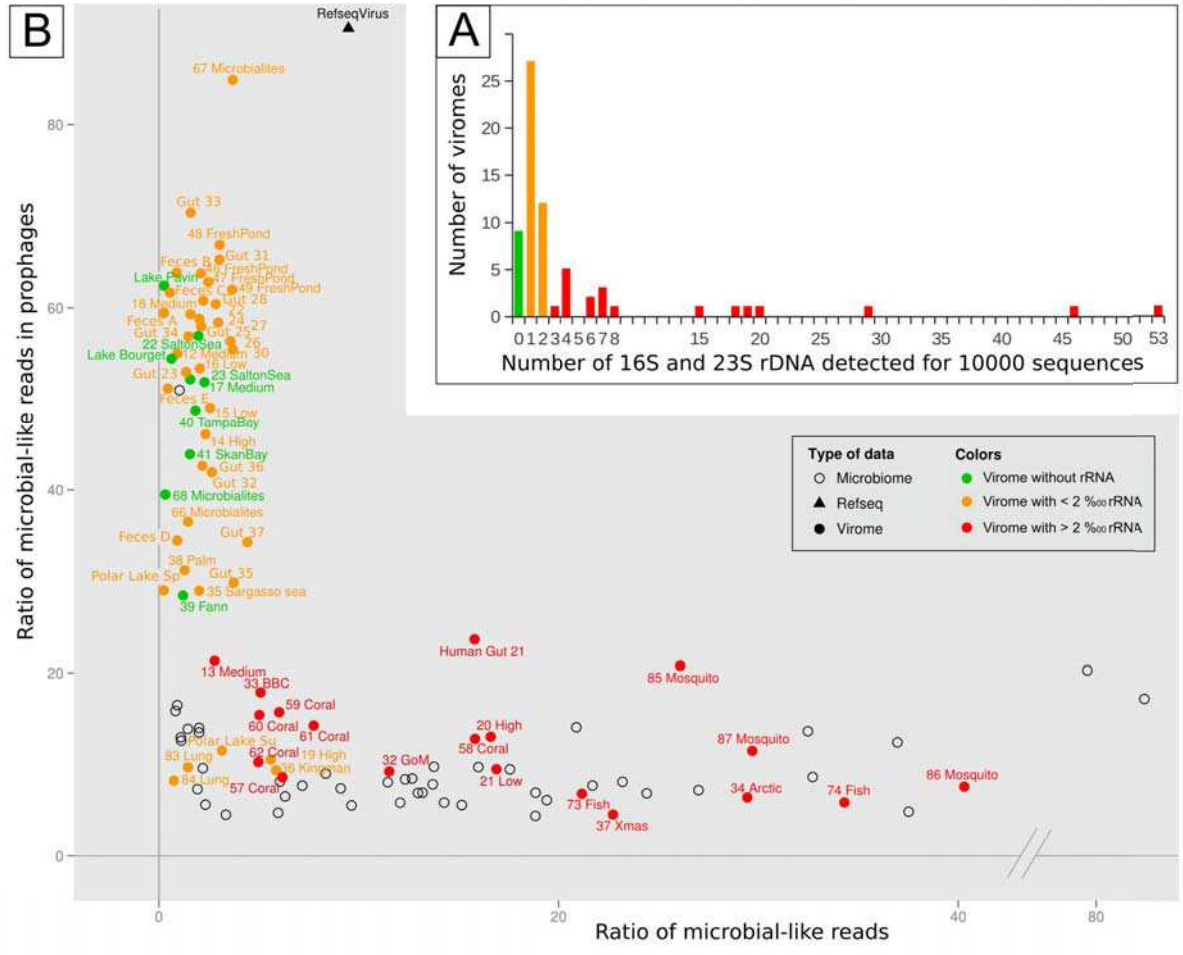


Figure 1: A. Distribution of relative number of rDNA genes detected in viromes. The three defined categories are colored: green for virome free from cellular DNA, orange for a low level of cellular DNA, and red for high level of cellular DNA. **B. PHR / MHR plot for each metagenome, either viral (filled dots, colored by their ecosystem types) or microbial (black circles).** For each dataset, the Microbial Hit Ratio (MHR) represents the proportion of reads having a significant similarity in a prokaryote genome. For reads having a hit in a bacterial genome, the Prophage Hit Ratio (PHR) represents the proportion of these microbial reads that are found in a prophage-like region. Viromes are colored according to their number of rDNA gene detected.

answers the question "how many of these matches are found in a prophage". Viromes PHR extended from 4.5 to 84.9 % (37.7 % on average, Fig. 1B, Table S2). The PHR was also computed for microbiomes (10.1 % on average) and, as expected, was very close to the proportion of prophage-like genes in microbial genomes (11%).

According to the PHR calculations, we formulated the following postulate: “the more cellular sequences in a virome, the closer its PHR is to those of microbial metagenomes”. We then considered simultaneously the MHR and the PHR, for microbiomes and for the viromes of the three groups defined on the basis of the rDNA ratios. The resulting plot confirmed different characteristics for these three groups of viromes (Fig. 1B, Table S2):

- Viromes devoid of rDNA (depicted in green in Fig. 1) are clearly distinct from microbial metagenomes: microbial-like sequences in these viromes are rare (low MHR, 1.3% on average), and most of them match the detected prophage-like regions (high PHR, 48.7% on average). These results further support the conclusion that these datasets can be considered as viromes *sensu stricto*.
- Viromes with low rDNA ratio (depicted in orange in Fig. 1) exhibit similar trends, most of them display low MHRs and high PHRs (average of 2.7 and 47.5 respectively). These results further indicate that most viromes in this category contains only a few (if any) microbial sequences.
- Viromes with the highest rDNA ratio (depicted in red in Fig. 1) are indistinguishable from microbial metagenomes. Indeed, the average MHR and PHR values for viromes in this category are very similar to those of microbiomes (MHR: 16.7% vs 15.8% and PHR: 12.4% vs 10.1% for these viromes and microbial metagenomes respectively), strongly indicating

that these viromes contains numerous microbial sequences.

As expected from the analysis of rDNA gene ratios, a gradient of presence of cellular DNA in viromes is confirmed by this analysis, with detected levels of microbial sequences ranging from quantitatively very low to high. The presence of cellular sequences in "red" viromes was also confirmed by the results of the recruitment plots as well as genome coverage ratio generated for selected virome-genome pairs (virome-genome pair with a low PHR; see Materials and Methods section). In "green" and "orange" viromes, the reads similar to non-prophage genes were often restricted to specific regions, and thus likely to be unpredicted prophage-like region and unknown genes shared by viruses and prokaryotes (Fig. 2A, Table S4). Conversely, all recruitment plots for "red" viromes displayed a hit distribution throughout the entire bacterial genomes with high gene coverage ratios (Fig. 2B, Table S4, all recruitment plots from Table S4 are available on http://metavir-meb.univ-bpclermont.fr/Recruitment_plots/recruitment_plot_gallery.php). A virome from Artic Sea samples (Angly *et al.* 2006) represents one of the most striking examples of a virome containing bacterial genomic DNA. Recruitment analysis showed that 91,315 reads from this virome can be matched to *Sphingopyxis alaskensis* (Fig. 2C), covering almost the entire genome. However, the route of acquisition of this bacterial DNA remains to be determined.

Source of prokaryotic sequences : GTA are invited to the party

The presence of prokaryotic DNA in viromes could to be due to intrinsic technical imperfections of virome preparation procedure (Forterre *et al.* 2013), which includes a combination of filtration, centrifugation, purification and extraction methods designed to select only viral capsids, and thus sequence only encapsidated genetic material (Vega Thurber *et al.* 2009). Indeed, even the most elaborate purification protocols are likely to be susceptible to residual contamination with microbial cells and/or free extracellular nucleic acids. Moreover, the virus-like particles isolated by these protocols (Breitbart *et al.* 2002; Vega Thurber et al. 2009) may include a non-negligible fraction of virus-like entities such as Gene Transfer Agents (GTAs) and DNA-containing membrane vesicles derived from cellular membranes (Kristensen *et*

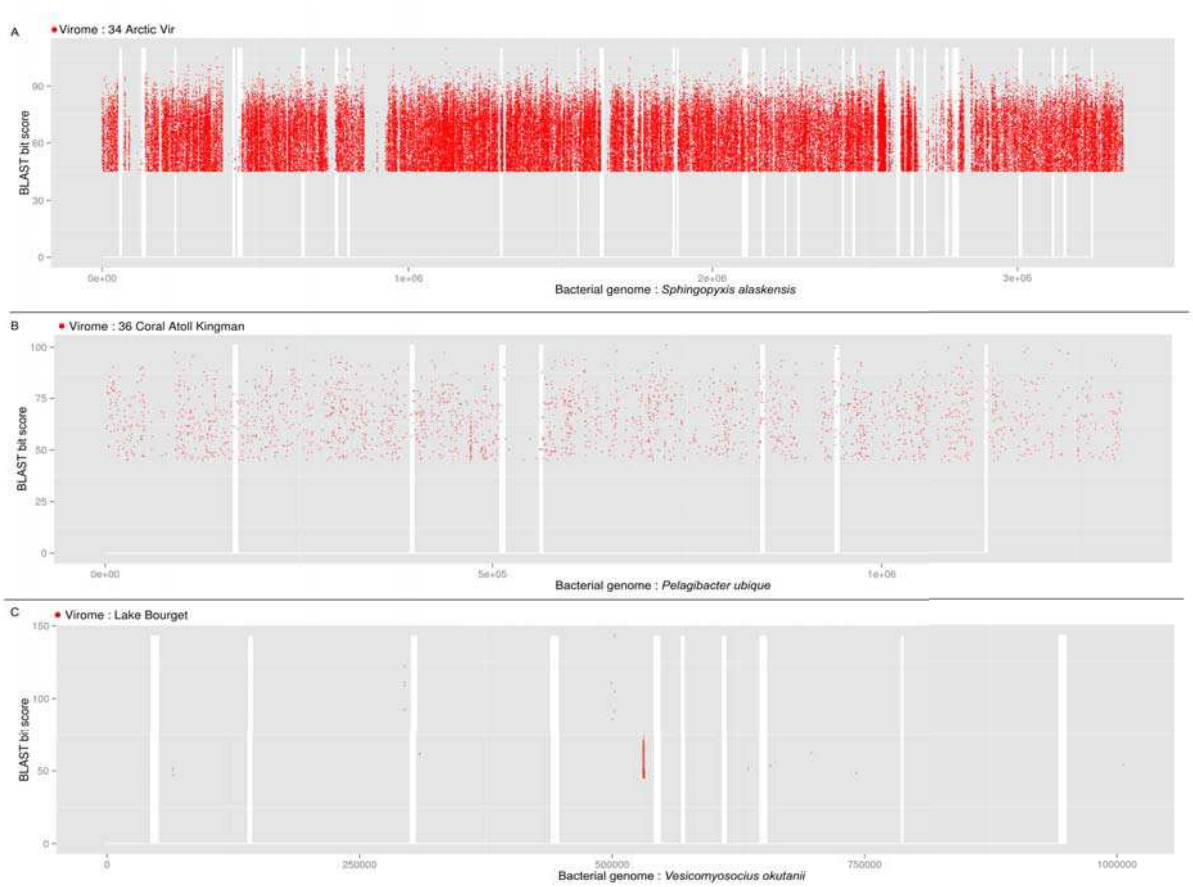


Figure 2: Recruitment plots for three virome-microbial genome associations. Virome reads were affiliated to the KEGG genome with the best tBLASTx score. Reads were then plotted at the position of the hit on the corresponding genome (x-axis), the sequence conservation being displayed as the identity percentage between read and genome on the y-axis. (A) Recruitment of 91,315 reads from the "34 Arctic Vir" virome by the genome of the Alphaproteobacteria Sphingopyxis alaskensis. (B) "36 Coral Atol" reads recruited by Pelagibacter ubique (1,973 reads) (C) 1,7444 reads of the Lake Bourget virome are recruited by Candidatus Vesicomyosocius okutanii.

al. 2010; Forterre *et al.* 2013). The GTAs are host-encoded virus-like elements that package random fragments of the host chromosome (McDaniel *et al.* 2010). Structurally, GTAs resemble small tailed phages (Yen *et al.* 1979), but do not possess any of the functions/properties (*e.g.* plaque formation, transmission of viral genes) that are typically associated with phages (Yen *et al.* 1979; Solioz *et al.* 1975). In our attempt to identify the origin of prokaryotic material in viromes, we verified the viability of the "GTA hypothesis" presented by Kristensen and co-authors (Kristensen *et al.* 2010**).**

For this purpose, each prokaryotic genome from KEGG database was analyzed for the presence of potential GTA gene clusters similar to the 4 GTA gene clusters reported previously (Table S3). We identified 72 prokaryotic strains (~6% of the known prokaryotic genomes), predominantly affiliated to the *α-proteobacteria*, containing putative GTA gene clusters (Table S3). We then identified, which of the 50 genomes detected in the "red" viromes (Table S4) exhibit GTA gene clusters. From this analysis, a dichotomous distribution of viromes emerged:

- **Eukaryote-associated samples** appear to be free from GTA as only ≈ 9% of the detected bacterial genomes exhibited GTA gene clusters.
- **Marine samples** could contain a significant amount of GTA particles, as > 50 % of the retrieved bacterial genomes contained at least one GTA cluster.

These two observations are consistent with previous results. First, virus-like particles (VLP) are described as difficult to purify from eukaryote-associated sample, due to the viscosity of the samples (Vega Thurber *et al.* 2009). We hypothesize that for certain systems (e.g., lung, mosquitoes, coral reef, etc.), it might be nearly impossible to obtain a pure viral material. However, without such studies, we would virtually have no information about these viral communities. Conversely, protocols and methods are well established for isolation of VLPs from aquatic samples. In the case of the seawater viromes, the presence of microbial sequences could thus be linked with the presence of GTA rather than technical limits. Our analysis showing a high ratio of GTA-encoding bacterial genomes in marine viromes is consistent with the high abundance of GTA particles predicted in marine bacterioplankton (McDaniel *et al.* 2010). Indeed, after their first discovery in the bacterium *Rhodobacter*, GTA were subsequently identified in many diverse prokaryotes and especially among *α-proteobacteria,* particularly in marine *Roseobacter* (Biers *et al.* 2008; Lang & Beatty 2007). Thus, GTA could be of major importance for directed gene transfer between phylogenetically related bacteria in low-density habitats such as seawater.

Toward a new picture of virus-associated functional profiles

The enrichment in virus-like particles in viromes does not result in significant differences in the functional profiles between viromes and microbiomes (Pearson correlation coefficient of 0.93, Fig. 3A). This is consistent with a previous observation (Dinsdale *et al.* 2008), and was suggested to be, at least partly, due to the registered functional categories in databases, which describe cellular rather than viral functions (Kristensen *et al.* 2010). The recently updated version of the SEED database, which offers a category entitled "Phages, prophages, transposable elements, plasmids", now allows testing this possibility. As expected, the latter category was more commonly found in viromes than in microbiomes (Fig. 3A); however, typical cellular functional categories remained abundant in viromes. Another explanation is that amount of prokaryotic DNA in viromes (through GTA and other non-viral entities) introduces a bias into functional profile analyses (Kristensen *et al.* 2010). Following identification of viromes with a high number of cellular-originating sequences, we postulated that a new picture of the functional profiles of viral communities might emerge from these data. To test this hypothesis, we computed functional profiles using viromes with clearly identified microbial-originating sequences ("red" viromes, Fig. 3B) on one side and viromes considered as mostly composed

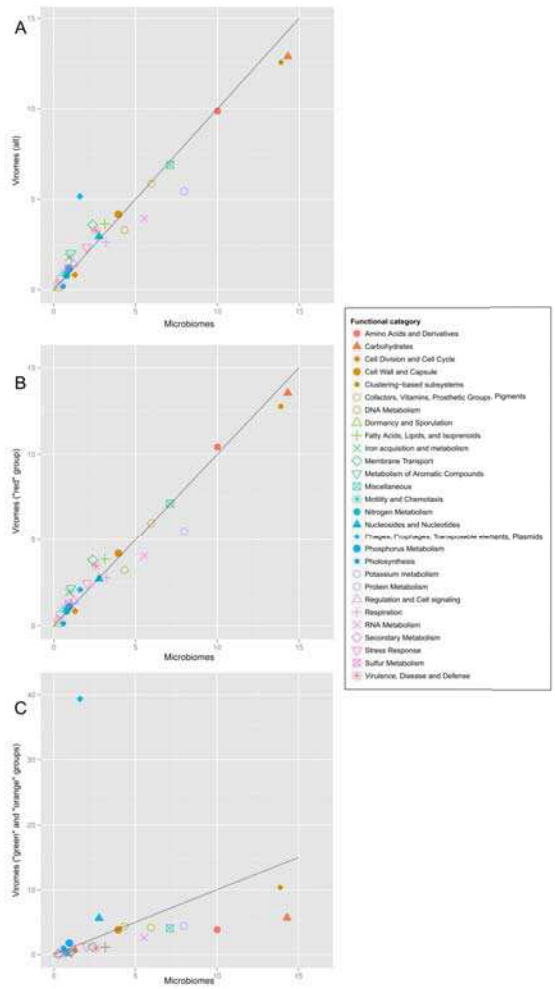


Figure 3: Comparison of the functional profiles of viromes and microbiomes, considering (A) all viromes, (B) viromes with clearly identified microbial-originating sequences ("red" viromes) and (C) viromes considered as mostly composed of viral sequences ("green" and "orange" viromes). The percentage of reads affiliated to each SEED category (level 1) is indicated for microbiomes (x-axis) and viromes (y-axis).

of viral sequences ("green and orange" viromes, Fig. 3C) on the other side. The functional profiles obtained for these two sets of viromes were very different (Fig. 3B and Fig. 3C). The functional profile of the first category of viromes was strongly correlated to the functional profile of microbiomes (Pearson correlation coefficient of 0.98), and the typical viral category "Phages, prophages, transposable elements, plasmids" ranked only at the 17th position in these viromes(2.09% of the functions, Fig. 3B). Conversely, a low correlation was found between functional profiles of the second category of viromes and microbiomes (Pearson correlation coefficient of 0.18), and these viromes displayed a strong enrichment in phage-like genes (39.8% for "Phages, prophages, transposable elements, plasmids"). Furthermore, prevalence of other categories in viromes and microbiomes was also no longer equivalent: these "green" and "orange" viromes were indeed depleted of typical cellular categories rarely observed in sequenced phages (*e.g.* "Cofactors, vitamins, prosthetic groups, pigments") but cellular categories commonly identified in known phages were retrieved (*e.g.* "Nucleosides and nucleotides", "DNA metabolism", Fig. 3C).

From this analysis, we demonstrated that the presence of bacterial DNA in several viromes biased the previous functional analyses of viromes leading to an artifactual correlation between functional profiles of viromes and microbiomes. Even if all functional categories are retrieved in "viral-only" viromes, indicating that all types of bacterial genes could be carried by the viral community, their proportions in viromes are highly different from those in microbiomes. Moreover, cellular sequences in viromes can have significant effects on the conclusions drawn from the functional analyses of these datasets. For example, the category "Motility and chemotaxis" enriched in viromes compared to microbiomes (1.00% and 0.66%) has been previously proposed as "an unexpected example of specialized metabolisms being carried within the viromes" (Dinsdale *et al.* 2008), but, according to our analysis, we postulate that this result was artifactual and linked to the presence of cellular DNA in viromes (enrichment of only 0.37 % for "green" and "orange" viromes, Fig. 3C).

Viral pan-genome encompass an unexpected diversity of metabolism genes

Due to its numerical vastness and genetic diversity, the virosphere is expected to embrace a tremendous functional potential. However, the extent of this potential remains unclear. Furthermore, the finding that a number of published viromes is also composed of cellular sequences non associated to a viral genome suggests that conclusions originally drawn from the analyses of the complete set of viromes might be inaccurate for depicting the functional potential of viruses *per se*. Obviously, the validity (and value) of the results is directly proportional to the "purity" of the analysed dataset. Even if this presumed slight presence of non-viral DNA (*i.e.* "orange" viromes) generates only a background noise in a large spectrum analysis such as functional profiling, it can bias the results when considering cellular functions one by one. Thus, in order to increase the likelihood of functional assignments being associated with viruses rather than cellular organisms, we only considered "green" viromes, in which no rDNA sequences were found.

A total number of 1,233 different KEGG orthology (KO) groups were detected in this dataset from the total of 14,645 KO groups present in KEGG database. Comparison of these 1,233 KO groups against the viral RefSeq sequences showed that 30% of them are represented in complete viral genomes. The most retrieved KO groups are often those already characterized: 75% of the highly-retrieved KO (associated to more than 20 virome sequences) are also represented in the complete viral genomes. The majority of these KO groups could be affiliated to dsDNA viruses infecting bacterial (*Myoviridae, Siphoviridae, Podoviridae*, *Tectiviridae*) and eukaryotic (*Phycodnaviridae*, *Mimiviridae*, *Iridoviridae*, "Marseilleviridae", *Nimaviridae*, *Baculoviridae*, *Nudivirus*) hosts. The KO groups present in the filtered dataset include proteins involved in all steps of viral infection cycle, i.e., virion morphogenesis (structural proteins, genome packaging enzymes), viral genome transcription, replication, recombination and repair as well as cell lysis (e.g., diverse peptidoglycan-digesting enzymes). These functional categories are well represented in the currently available viral genomes and will not be further discussed. Perhaps more unexpected was identification of diverse protein functions responsible for modulation of cellular metabolism and virus-host interactions. Below, we briefly outline the most prominent KEGG functional categories retrieved and highlight potential roles of these proteins in the framework of viral infection cycles.

Energy metabolism genes.

One of the landmark discoveries of the past decade was the identification of functional photosystem (PS) II genes in the genomes of cyanophages (Lindell *et al.* 2004). More recently, analysis of metagenomic data revealed that marine cyanophages might also encode the entire suite of proteins composing PSI (7 proteins) (Sharon *et al.* 2009). These findings have clearly demonstrated that viruses may play an active role in energy transformation. In accordance with previous results, our list of KO groups included components of both PSII (including proteins D1 and D2) and PSI (including PsaA and PsaB, see Table 1 and Figure 6). These photosynthesis genes did not present the same pattern of distribution: PSI genes were found exclusively in marine viromes, while those of PSII were also present in freshwater and hypersaline environments (Table S5). Surprisingly, our analysis suggests that besides photosynthesis genes viruses may encode a set of proteins involved in oxidative phosphorylation. We identified several components of the prokaryotic electron transport chain *Complexes I*, *II, III* and *IV* (Table 1, Figure 6). Intriguingly, it appears that viruses might also harbour genes for at least some subunits (α,b) of the *F_0F_1 ATP synthase* (also referred to as Complex V) as well as genes for inorganic *pyrophosphatase* (Ppa), which is responsible for supplying inorganic phosphate for the ATP synthesis by ATP synthase. Notably, the latter set of enzymes might also operate in conjunction with the photosystem genes. Indeed, genes for the *a*, *b* and *c* subunits of the *F_0F_1 ATP*

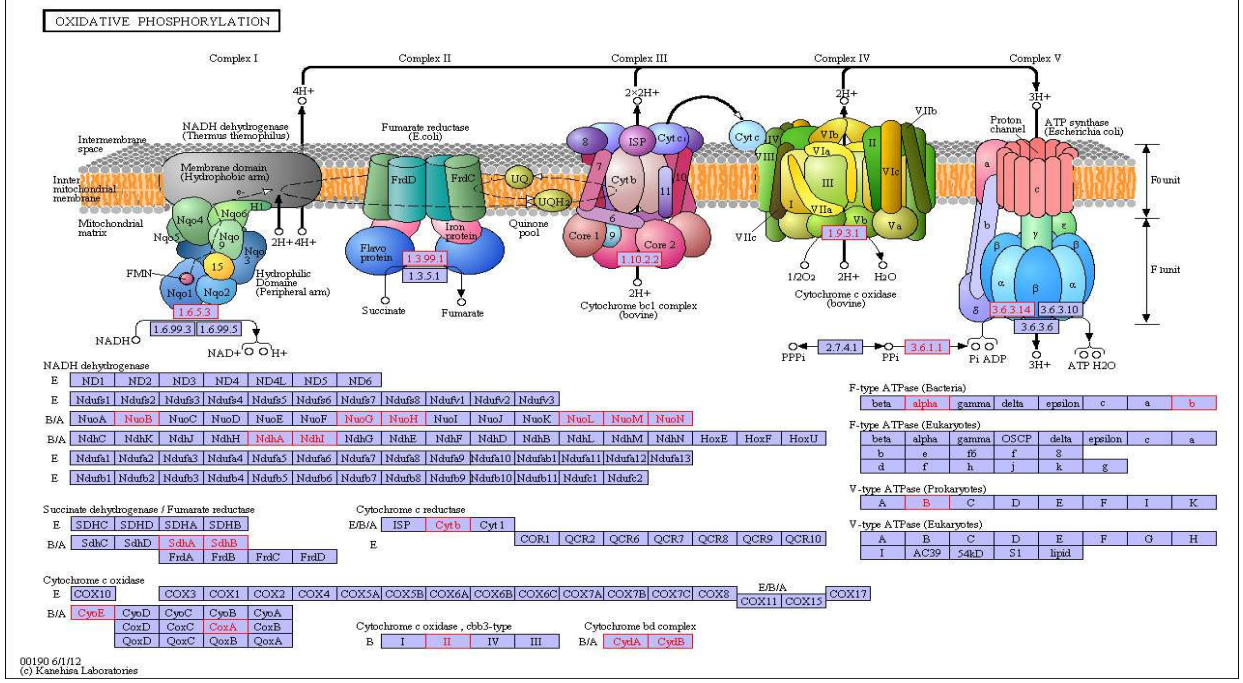


Figure 4: Mapping of virome-retrieved functions on oxydative phosphorylation pathway. On this general representation of the oxydative phosphorylation pathway, KO retrieved in uncontaminated viromes are highlighted in gray on the list of KO at the bottom, and when possible on the chart at the top.

synthase have been recently reported in the environmental GOS cyanophage clone JCVI_SCAF_1096628171668 (Philosof *et al.* 2011). Similarly, metagenomic studies have previously suggested that cyanophages might harbour the ndhI, ndhD and ndhP genes of the Complex I (Philosof et al. 2011; Sharon et al. 2009). Finally, we found both subunits (CydA and CydB) of the two-component *cytochrome bd quinol oxidase*, which is associated with microaerobic dioxygen respiration (Poole & Cook 2000).

Carbon metabolism genes.

Unexpectedly, the dataset contained a substantial number of enzymes involved in such fundamental cellular metabolism pathways as glycolysis, tricarboxylic acid (TCA) cycle and pentose phosphate pathway (Table 1, Table S5). With few exceptions, genes of this category are not typically found in viral genomes.

Glycolysis. Glycolysis is a universal metabolic pathway of converting glucose into pyruvate and generating small amounts of the high-energy compounds ATP and NADH. The glycolytic breakdown of glucose in anaerobic or severely-hypoxic conditions is the sole source of ATP for many microorganisms. We identified 11 KO groups that were related to glycolysis pathway and detected more than once in viromes (Table 1, Table S5). A growing body of evidence suggests that viruses might modulate the host metabolism according to their needs. For example, it has been suggested that cyanophage-encoded proteins may modify the photosynthetic electron transfer chain such that the cyclic electron flow around PSI would be favoured over the linear one, leading to preferential production of ATP (Philosof *et al.* 2011). In this light, it is tempting to speculate that the viral versions of glycolysis enzymes might be differentially susceptible to allosteric regulation compared to their cellular counterparts as to maximize the energy production for optimal virus replication.

Tricarboxylic acid cycle (TCA) and pyruvate metabolism. In aerobic conditions, glycolysis, fat and protein catabolic pathways converge on the TCA cycle. As a result, carbohydrates, fatty acids and amino acids are oxidized to CO_2 with most of the energy of oxidation temporarily held in the electron carriers $FADH_2$ and NADH, which eventually enter the respiratory chain where the energy of electron flow is converted to ATP. Thus, TCA cycle represents the central catabolic pathway in aerobic organisms. We identified 10 non-singleton virome-associated KO groups involved in TCA cycle (KO groups detected more than once in viromes), including *pyruvate dehydrogenase* (E1 subunit α and β), which is responsible for converting pyruvate generated during glycolysis into acetyl-CoA. In addition, 11 non-singleton KO groups were found to be affiliated with the pyruvate metabolism pathway (ko00620) (Table 1, Table S5).

Pentose phosphate pathway (PPP). Ten non-singletons KO groups in our dataset mapped to the PPP, which represents an alternative route of glucose metabolism. PPP is a two-phase pathway leading to production of reducing equivalent NADPH (during oxidative phase) and pentose phosphates for synthesis of nucleotides and amino acids (during non-oxidative phase). It has been previously demonstrated that some cyanophages encode functional homologues of cyanobacterial *transaldolase* (TalC)

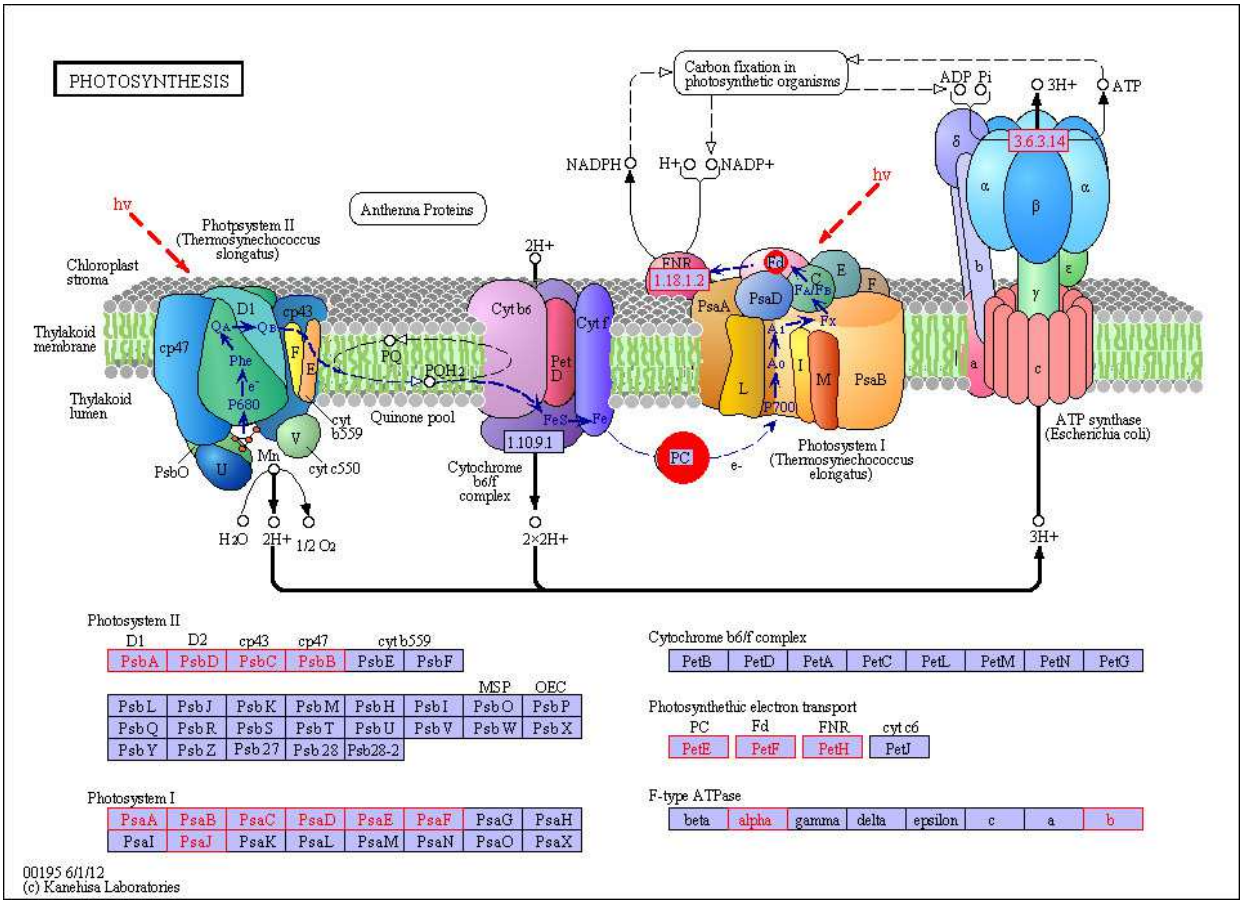


Figure 5: Mapping of virome-retrieved functions on the different types of photosystem. On this general representation of the photosystems, KO retrieved in uncontaminated viromes are highlighted in gray on the list of KO at the bottom, and when possible on the chart at the top.

(Thompson et al. 2011; Sullivan et al. 2005), *6-phosphogluconate dehydrogenase* (Gnd) and *glucose-6-phosphate 1-dehydrogenase* (G6PD) (Sullivan et al. 2010), key enzymes of the PPP. TalC, Gnd and G6PD were all retrieved in our analysis among high-confidence virome-associated KO groups. In addition to the three enzymes mentioned above, our data suggests that viruses carry genes for other PPP enzymes, including *transketolase* (Tkt), *ribose-phosphate pyrophosphokinase* (PRPS), *ribose-5-phosphate isomerase* (rpiB) and *fructose-biphosphate aldolase of class I* and *II* (fbaB and fbaA ; Table 1 , Table S5**)**. Notably, G6PD catalyses the first, essentially irreversible reaction in the oxidative phase of the PPP and is the rate-limiting enzyme of the pathway. Expression of the viral G6PD might thus stimulate PPP indicating that this pathway is beneficial for virus replication. Indeed, it has been shown that cyanophages specifically direct carbon flux away from the Calvin cycle toward the PPP, this way ensuring that the ATP and NADPH produced by photosynthesis are not consumed in the Calvin cycle but are rather used to fuel phage dNTP biosynthesis (Thompson et al. 2011). This is consistent with the identification of the virome-associated genes for PRPS, one of the key enzymes in the *de novo* and *salvage* biosynthesis of nucleotides.

Glycan biosynthesis and modification genes.

Genes of this functional category are frequently found in the genomes of eukaryotic DNA viruses, particularly in members of the *Phycodnaviridae*, which encode most of the machinery to glycosylate their glycoproteins (Van Etten et al. 2010). Similarly, many archaeal viruses possess genes for diverse glycosyltransferases (Krupovic et al. 2012), while some encode sugar modification enzymes (e.g., GDP-mannose 4,6-dehydratase in *Aeropyrum* coil-shaped virus) (Mochizuki et al. 2012). Consistently, it has been demonstrated that virion proteins of many eukaryotic and archaeal viruses are glycosylated. Bacterial T4-like viruses, on the other hand, glucosylate certain bases of their genomic DNA, rather than virion proteins (Moréra et al. 1999). Moreover, phycodnavirus PBCV-1 encodes three enzymes responsible for synthesis of the extracellular matrix polysaccharide hyaluronan (Graves et al. 1999). Thus, virus-encoded carbohydrate metabolism and glycan biosynthesis genes may be responsible for modulating various aspects of virus-host interactions, such as receptor recognition, evasion of immune system, resistance to proteases, etc. Consistent with their importance in viral reproduction cycles, we retrieved virome-associated KO groups responsible for sugar metabolism and glycan biosynthesis. Besides enzymes already observed in viruses this set also included glycan biosynthesis enzymes that so far have been only

associated with cellular organisms, indicating that virus-encoded carbohydrate modification machinery might be more versatile than currently appreciated (Table 1, Table S5).

Amino acid metabolism and transport genes. Genomes of dsDNA viruses sporadically contain genes for enzymes involved in amino acid biosynthesis and metabolism. Although genes belonging to this functional category are most often found in eukaryotic large DNA viruses (e.g., *Mimiviridae* and *Phycodnaviridae*), they are also occasionally present in moderately sized genomes of bacterial viruses (Villion et al. 2009). We have found a number of new virome-associated KO groups responsible for amino acid metabolism (Table 1, Table S5). Notably, all pathways involving metabolism of proteinogenic amino acids were represented in the "uncontaminated" viromes, albeit to variable extents. In addition, we identified a full suite of genes (livG, livF, livH, livM and livK) constituting an ABC transporter specializing in the uptake of branched amino acids. Indeed, putative amino acid transporters have been previously found to be encoded by large eukaryotic DNA viruses (Fischer et al. 2010), but also by certain much smaller viruses infecting prokaryotes (Prangishvili et al. 2013). The exact role of viral proteins involved in amino acid transport and metabolism has not been studied in the framework of the viral infection cycle. However, it is possible that expression of the corresponding viral enzymes optimizes the intracellular balance of amino acids, which might be altered because of viral infections (Hortin et al. 1994; Agudelo-Romero et al. 2008).

Translation genes.

Sequencing of the Mimivirus genome revealed that viruses might occasionally encode proteins involved in translation, such as aminoacyl tRNA synthetases (aaRS) and translation initiation and elongation factors (Raoult *et al.* 2004). This finding has been subsequently confirmed by additional genome sequences of large eukaryotic (Arslan *et al.* 2011; Fischer *et al.* 2010) and, more recently, bacterial (Hendrix 2009) viruses. To date, members of the *Mimiviridae* were found to encode seven different aaRS – ArgRS, TyrRS, CysRS, MetRS, IleRS, TrpRS, AsnRS (Arslan et al. 2011; Fischer et al. 2010; Raoult et al. 2004), while *Bacillus megaterium* phage G carries a gene for SerRS (Hendrix 2009). In the uncontaminated viromes, we identified aaRS genes specific for 18 of the 20 proteinogenic amino acids as well as several genes for enzymes involved in modification of aminoacyl-tRNAs, including methionyl-tRNA formyltransferase (required for formation of formylMet-tRNA, an initiator tRNA in bacteria, mitochondria and chloroplasts) and aminoacyl-tRNA amidotransferase (Table 1, Table S5). In addition, we found genes for translation initiation (IF-1, 2, and 3), elongation (EF-G) and peptide chain release (RF-1 and RF-3) factors.

As expected, no rRNA genes were retrieved. However, several rRNA modification enzymes, such as rRNA methyltransferases and rRNA pseudouridine synthase, were identified. Finally, a set of 6 non-singleton ribosomal proteins were also present in the filtered dataset (Table 1, Table S5). Currently, there are no precedents of ribosomal proteins being encoded by viruses. Thus, it is not clear whether the two genes signify the presence of cellular sequences or genuine gene acquisitions by viruses. However, a point can be made that there is no obvious reason why these ribosomal protein genes, which are detected up to 18 times within 4 different viromes, should be recovered in the viral fraction to the exclusion of all other ribosomal genes, including those for rRNA that are often present in multiple copies per cellular genome and are statistically more likely to be identified among cellular-originating sequence (Klappenbach *et al.* 2001). Ribosomal protein genes are known to be transferred horizontally (Brochier *et al.* 2000; Makarova *et al.* 2001; Garcia-Vallvé *et al.* 2002; Coenye & Vandamme 2005), although the particular routes of such transfer remain unclear. One possibility, which might be strengthened by observations presented above, is that viruses serve as vehicles for horizontal transfer of ribosomal protein genes, as is the case with many other cellular genes (Krupovic *et al.* 2011). What could be a role of ribosomal protein in the course of a viral cycle? Modification of the ribosomes by viral versions of the ribosomal proteins might allow viruses to overcome a translational shutoff in the host, which may be potentially triggered by viral infection. Indeed, bacterial viruses are known to induce the toxin components of certain toxin-antitoxin systems (Hazan & Engelberg-Kulka 2004), some of which are known to poison or stall the ribosomes (Liu *et al.* 2008). Alternatively, many ribosomal proteins perform extraribosomal functions, a phenomenon known as moonlighting (Aseev & Boni 2011; Copley 2012) . Notably, protein S1, one of the most detected in our dataset, is one of such proteins; in addition to being a structural component of the ribosomes, S1 regulates expression of several ribosomal operons, including that of its own (Aseev & Boni 2011). Finally, Qβ and other leviviruses hijack S1 to serve as a subunit of their RNA replicases (Wahba *et al.* 1974). It is thus possible that viruses recruit ribosomal protein genes for functions that have little to do with ribosome structure.

Peculiarly, ribosomes represent one of the final frontiers distinguishing viruses and cellular organisms (Raoult & Forterre 2008), at least from the genomic perspective. Additional efforts focused on exploration of genetic diversity in the virosphere and especially these intriguing ribosomal proteins are undoubtedly needed to resolve this puzzle.

Conclusions

The putative presence of non-viral sequences in viromes undoubtedly raises questions about these datasets, but must not be seen as challenging all previous results and conclusions. Indeed, the presence of cellular DNA in viromes has certainly little effect on the analysis and interpretation of sequences that can be unequivocally assigned to viruses (*i.e.* when reasonably close homologues are present in the genomes of cultivated viruses), as was the case in most virome studies published to date. However, questions related to functional capacity of uncultured viral communities, and specifically the diversity of microbial-like genes in viral genomes, require

all sequences in the viromes to be of viral origin in order to be rigorously addressed. If the latter point is neglected, the validity and value of conclusions drawn from the virome analyses become questionable, as illustrated by the results presented in this study.

Our study also pinpoints the differential source of cellular sequences in viromes obtained from different environments, stressing out the role of GTA and other DNA-containing membrane vesicles in the case of seawater samples. Unfortunately, as GTA display a viral capsid structure, it is likely that no preparation step will be able to separate them from actual viral capsids, hence this type of "contamination" is probably irremediable. In such case, downstream bioinformatics analysis will be needed to check their presence in viromes.

Eventually, one of the most significant findings resulting from this analysis was the abundance and global distribution of virome-associated operational (metabolic) genes. Indeed, it appears that in all analysed biomes, viruses intensively tinker with the metabolism of their hosts. A great deal of functional and genomic data on photosynthetic genes in cyanophages made this viral group stand out as an exception, or a peculiarity within the virosphere in the eyes of many (micro)biologists. Here we provided evidence suggesting that beside photosynthesis, viruses might tap into such central metabolic pathways as oxidative phosphorylation, glycolysis, tricarboxylic acid cycle and pentose phosphate pathway. It is noteworthy that some of these metabolic enzymes have been previously identified in viral genomes, Although the available scattered data did not allow to draw generalizing conclusions on the role of viruses in the cellular metabolism beyond particular virus-host systems, our analysis of viromes issued from diverse environments illuminates a somewhat unexpected picture of global "viral" metabolism, suggesting that viruses might actively dictate the metabolism of infected cells on a global scale.

Materials and methods

Genomic and metagenomic sequence data

The prokaryotic sequences used as references (1312 complete genomes and the corresponding 4,457,923 protein sequences) originated from KEGG database (Kanehisa *et al.* 2012). Viral genomes (2,852) and the encoded protein sequences (104,703) were obtained from RefSeqVirus database (Pruitt *et al.* 2007). Reference databases were downloaded in June 2011 and March 2012 respectively. The metagenomic data were composed of 45 microbial and 67 viral publicly available metagenomes (Dinsdale *et al.* 2008; López-Bueno *et al.* 2009; Roux *et al.* 2012; Minot *et al.* 2011; Kim *et al.* 2011; see Table S1).

Detection of ribosomal rDNA in viromes

Genes encoding the 16S and 23S rRNAs (from prokaryotic genomes) were identified in viromes using rna_hmm, a sensitive tool based on HMM search (Huang *et al.* 2009). rDNA gene prediction were then checked through a BLAST comparison to the SILVA database (Quast *et al.* 2013).

Detection of prophage-like regions in prokaryotic genomes

Prokaryotes sequences similar to viral sequences, referred as viral-like-genes, were identified by BLASTp comparison (Altschul *et al.* 1990) according to bit-score and E-value thresholds of 50 and 0.001 respectively. Prophage-like regions were then defined according to the following criteria: a region of 4 or more genes, containing at least 1 viral-like gene, and composed of only viral-like genes or hypothetical protein-coding genes (*i.e.* bacterial genes for which no function are identified, noted by the keywords "hypothetical protein" or "putative protein" in their annotation). Although several more sophisticated prophage detection tools are available (Akhter *et al.* 2012; Lima-Mendez *et al.* 2008), we intentionally relied on such "naive" prophage definition criteria in order to detect not only functional prophages but also the defective and degenerated ones.

Comparison of viromes and microbiomes to prokaryotic genomes

To avoid bias resulting from differences in the length of metagenomic sequences (Table S1), all viromes reads were randomly truncated to 100 bp before proceeding to comparison. Viral genomes from RefSeqVirus were also truncated to 100 bp and used as a simulated metagenome (100,000 sequences of 100 bp generated with Grinder (Angly *et al.* 2012)). All resulting 100 bp reads were compared to prokaryotic genomes using tBLASTx (bit-score and E-value thresholds of 50 and 0.001, respectively). Each read was affiliated to its best-matched prokaryotic genome enabling to determine, for each metagenome, the Microbial Hit Ratio (MHR):

MHR = (Number of reads with a hit in KEGG database/ Total number of reads) x 100

According to the prophage-like regions identified in Step 1, the Prophage Hit Ratio (PHR) was determined:

PHR = (Number of reads with a prophage as best hit in KEGG database/ Number of reads with a hit in KEGG database) x 100

These two ratios are summarized on the plot PHR versus MHR.

Identification of the origin of cellular DNA in viromes

To ensure that a low PHR did not result from affiliation of reads to specific genomic regions, such as unknown viral genes or isolated genes common to prokaryotes and viruses, a complementary procedure was performed. For each virome, recruitment plots were generated for each genome recruiting 500 or more reads. Plots were manually inspected when the prophage-ratio of a virome-genome pair was lower than the prophage-ratio of the genome (+ 5 %). This detailed analysis of virome-genome pairs enabled us to identify the genome(s) involved for each virome in which cellular DNA was detected. All recruitment plots are available on a dedicated web page : http://metavir-meb.univ-bp.clermont.fr/Recruitment_plots/recruitment_plot_gallery.php

Proportion of genomes containing GTA gene

clusters in viromes

To determine the possible presence of cellular DNA in viromes due to Gene Transfer Agents (GTA), 4 previously described GTA gene clusters (Table S3) were used to detect potential homologous clusters in prokaryotic genomes using BLASTp (bit-score and E-value thresholds of 50 and 0.001, respectively). Three of these clusters are well documented and represent experimentally confirmed GTA gene clusters (Lang *et al.* 2012): one in the Spirochaetes *Brachyspira hyodysenteriae* (Matson *et al.* 2005) and two in the α-proteobacteria *Rhodobacter capsulatus* (Lang & Beatty 2007) and *Silicibacter pomeroyi* (Biers *et al.* 2008). The fourth cluster used is a predicted GTA-encoding genomic region from *Methanococcus voltae* A3 (Krupovic et al. 2010), a methanogenic, anaerobic archaeon previously demonstrated to produce GTA particles (Eiserling et al. 1999). Genomic regions enriched in GTA-like genes were manually inspected and GTA clusters in the reference set of prokaryotic genomes were predicted according to the following conditions: the absence of gene coding for an integrase, the size of the genomic region considered (< 40 genes) and the genomic neighborhood of the putative cluster. According to the identification of cellular DNA in viromes (Step 3) and of genomes containing GTA gene clusters, the ratio of GTA-containing genome was calculated for each virome.

Functional analysis of viromes and microbiomes

- Functional profiles of 42 viromes and of 45 microbiomes, previously analyzed by Dinsdale and collaborators (Dinsdale *et al.* 2008) and Kristensen and collaborators (Kristensen *et al.* 2010) (Table S1) were downloaded from the Mg-Rast web-server (Meyer *et al.* 2008) and were compared. Three comparisons were performed: all viromes *vs* all microbiomes, viromes of Group A *vs* all microbiomes, viromes of Group B *vs* all microbiomes. Plots were generated for each combination and Pearson's correlation coefficients were computed.

- Functional annotation of the 9 viral-only viromes was performed using tBLASTx comparison between viromes and the KEGG database (Step 5), KEGG Orthology (KO) system and the associated online pathway representation (Aoki-Kinoshita & Kanehisa 2007).

Author contributions

SR, DD and FE designed the experiment, SR and FE performed the experiment, SR, MK, PF and FE analyzed the results, SR, MK, DD, PF and FE wrote the manuscript

Acknowledgments

SR was supported by a PhD grant from the French defense procurement agency (DGA, Direction Générale de l'Armement).

References

Agudelo-Romero P, Carbonell P, De la Iglesia F, Carrera J, Rodrigo G, Jaramillo A, *et al.* (2008). Changes in the gene expression profile of Arabidopsis thaliana after infection with Tobacco etch virus. Virology journal 5:92.

Akhter S, Aziz RK, Edwards RA. (2012). PhiSpy: a novel algorithm for finding prophages in bacterial genomes that combines similarity- and composition-based strategies. Nucleic acids research 40:1–13.

Allen MJ, Wilson WH. (2008). Aquatic virus diversity accessed through omic techniques: a route map to function. Current opinion in microbiology 11:226–32.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ, Altschul, S F; Gish, W; Miller, W; Wyeres, E W; Lipman DJ. (1990). Basic local alignment search tool. Journal of molecular biology 215:403–410.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C, *et al.* (2006). The marine viromes of four oceanic regions. PLoS biology 4:e368.

Angly FE, Willner D, Rohwer F, Hugenholtz P, Tyson GW. (2012). Grinder: a versatile amplicon and shotgun sequence simulator. Nucleic acids research 40:e94.

Aoki-Kinoshita KF, Kanehisa M. (2007). Gene annotation and pathway mapping in KEGG. In: Bergman, NH, (ed). Methods in molecular biology Vol. 396, Clifton, N.J., pp. 71–91.

Arslan D, Legendre M, Seltzer V, Abergel C, Claverie J-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proceedings of the National Academy of Sciences of the United States of America 108:17486–91.

Aseev L V., Boni I V. (2011). Extraribosomal functions of bacterial ribosomal proteins. Molecular Biology 45:739–750.

Bergh O, Børsheim KY, Bratbak G, Heldal M. (1989). High abundance of viruses found in aquatic environments. Nature 340:467–468.

Biers EJ, Wang K, Pennington C, Belas R, Chen F, Moran MA. (2008). Occurrence and expression of gene transfer agent genes in marine bacterioplankton. Applied and environmental microbiology 74:2933–9.

Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D, *et al.* (2002). Genomic analysis of uncultured marine viral communities. Proceedings of the National Academy of Sciences of the United States of America 99:14250–5.

Brochier C, Hervé P, Moreira D. (2000). The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. Trends in genetics : TIG 16:529–533.

Casjens S. (2003). Prophages and bacterial genomics: what have we learned so far? Molecular Microbiology 49:277–300.

Coenye T, Vandamme P. (2005). Organisation of the S10, spc and alpha ribosomal protein gene clusters in prokaryotic genomes. FEMS microbiology letters 242:117–26.

Copley SD. (2012). Moonlighting is mainstream: paradigm adjustment required. BioEssays : news and reviews in molecular, cellular and developmental biology 34:578–88.

Dimijian GG. (2000). Pathogens and parasites: strategies and challenges. Proceedings (Baylor University. Medical Center) 13:19–29.

Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, *et al.* (2008). Functional metagenomic profiling of nine biomes. Nature 452:629–32.

Edwards RA, Rohwer F. (2005). Viral metagenomics. Nature Reviews Microbiology 3:504–510.

Eiserling F, Pushkin A, Gingery M, Bertani G. (1999). Bacteriophage-like particles associated with the gene transfer agent of Methanococcus voltae PS. Journal of general virology 80:3305–3308.

Van Etten JL, Gurnon JR, Yanai-Balser GM, Dunigan DD, Graves M V. (2010). Chlorella viruses encode most, if not all, of the machinery to glycosylate their glycoproteins independent of the endoplasmic reticulum and Golgi. Biochimica et biophysica acta 1800:152–9.

Fischer MG, Allen MJ, Wilson WH, Suttle CA. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton. 107:1–6.

Forterre P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. Virus research 117:5–16.

Forterre P, Soler N, Krupovic M, Marguet E, Ackermann H-W. (2013). Fake virus particles generated by fluorescence microscopy. Trends in Microbiology 21:1–5.

Garcia-Vallvé S, Simó FX, Montero M a, Arola L, Romeu A. (2002). Simultaneous horizontal gene transfer of a gene coding for ribosomal protein l27 and operational genes in Arthrobacter sp. Journal of molecular evolution 55:632–7.

Graves M V, Burbank DE, Roth R, Heuser J, DeAngelis PL, Van Etten JL. (1999). Hyaluronan synthesis in virus PBCV-1-infected chlorella-like green algae. Virology 257:15–23.

Hazan R, Engelberg-Kulka H. (2004). Escherichia coli mazEF-mediated cell death as a defense mechanism that inhibits the spread of phage P1.

Molecular genetics and genomics : MGG 272:227–34.
Hendrix RW. (2009). Jumbo bacteriophages. Current topics in microbiology and immunology 328:229–40.
Hortin GL, Landt M, Powderly WG. (1994). Changes in plasma amino acid concentrations in response to HIV-1 infection. Clinical chemistry 40:785–9.
Huang Y, Gilna P, Li W. (2009). Identification of ribosomal RNA genes in metagenomic fragments. Bioinformatics 25:1338–40.
Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. Nucleic acids research 40:D109–14.
Kim M-S, Park E-J, Roh SW, Bae J-W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. Applied and environmental microbiology 77:8062–8070.
Klappenbach J a, Saxman PR, Cole JR, Schmidt TM. (2001). rrndb: the Ribosomal RNA Operon Copy Number Database. Nucleic acids research 29:181–4.
Kristensen DM, Mushegian AR, Dolja V V, Koonin E V. (2010). New dimensions of the virus world discovered through metagenomics. Trends in microbiology 18:11–19.
Krupovic M, Forterre P, Bamford DH. (2010). Comparative analysis of the mosaic genomes of tailed archaeal viruses and proviruses suggests common themes for virion architecture and assembly with tailed viruses of bacteria. Journal of molecular biology 397:144–60.
Krupovic M, Prangishvili D, Hendrix RW, Bamford DH. (2011). Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiology and molecular biology reviews : MMBR 75:610–35.
Krupovic M, White MF, Forterre P, Prangishvili D. (2012). Postcards from the edge: structural genomics of archaeal viruses. In: Łobocka, M & Szybalski, WT (ed). Advances in virus research, Vol. 82. Elsevier Inc., pp. 33–62.
Lang AS, Beatty JT. (2007). Importance of widespread gene transfer agent genes in alpha-proteobacteria. Trends in microbiology 15:54–62.
Lang AS, Zhaxybayeva O, Beatty JT. (2012). Gene transfer agents: phage-like elements of genetic exchange. Nature reviews. Microbiology 10:472–82.
Lima-Mendez G, Van Helden J, Toussaint A, Leplae R. (2008). Prophinder: a computational tool for prophage prediction in prokaryotic genomes. Bioinformatics 24:863–5.
Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. Proceedings of the National Academy of Sciences of the United States of America 101:11013–8.
Liu M, Zhang Y, Inouye M, Woychik NA. (2008). Bacterial addiction module toxin Doc inhibits translation elongation through its association with the 30S ribosomal subunit. 105:5885–5890.
López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, Alcamí A. (2009). High diversity of the viral community from an Antarctic lake. Science 326:858–61.
López-García P, Moreira D. (2009). Yet viruses cannot be included in the tree of life. Nature Reviews Microbiology 7:615–617.
Ludmir EB, Enquist LW. (2009). Viral genomes are part of the phylogenetic tree of life. Nature Reviews Microbiology 7:615–615.
Makarova KS, Ponomarev V a, Koonin E V. (2001). Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. Genome biology 2:0033.1–0033.14.
Matson EG, Thompson MG, Humphrey SB, Zuerner RL, Stanton TB. (2005). Identification of Genes of VSH-1, a Prophage-Like Gene Transfer Agent of Brachyspira hyodysenteriae. Journal of bacteriology 187:5885–5892.
McDaniel LD, Young E, Delaney J, Ruhnau F, Ritchie KB, Paul JH. (2010). High frequency of horizontal gene transfer in the oceans. Science 330:50.
Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, *et al.* (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics 9:386.
Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, Wu GD, *et al.* (2011). The human gut virome: inter-individual variation and dynamic response to diet. Genome Research 21:1616–1625.
Mochizuki T, Krupovic M, Pehau-Arnaudet G, Sako Y, Forterre P, Prangishvili D. (2012). Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. Proceedings of the National Academy of Sciences 109:1–6.
Monier A, Pagarete A, De Vargas C, Allen MJ, Read B, Claverie J-M, *et al.* (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. Genome research 19:1441–9.
Moreau H, Piganeau G, Desdevises Y, Cooke R, Derelle E, Grimsley N. (2010). Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. Journal of virology 84:12555–63.
Moréra S, Imberty a, Aschke-Sonnenborn U, Rüger W, Freemont PS. (1999). T4 phage beta-glucosyltransferase: substrate binding and proposed catalytic mechanism. Journal of molecular biology 292:717–30.
Ng TFF, Willner DL, Lim YW, Schmieder R, Chau B, Nilsson C, *et al.* (2011). Broad surveys of DNA viral diversity obtained through viral metagenomics of mosquitoes. PLoS One 6:e20579.
Philosof A, Battchikova N, Aro E-M, Béjà O. (2011). Marine cyanophages: tinkering with the electron transport chain. The ISME journal 5:1568–70.
Poole RK, Cook GM. (2000). Redundancy of aerobic respiratory chains in bacteria? Routes, reasons and regulation. Advances in microbial physiology 43:165–224.
Prangishvili D, Koonin E V, Krupovic M. (2013). Genomics and biology of Rudiviruses, a model for the study of virus-host interactions in Archaea. Biochemical Society transactions 41:443–50.
Pruitt KD, Tatusova T, Maglott DR. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic acids research 35:D61–5.
Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, *et al.* (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic acids research 41:D590–6.
Raoult D, Audic S, Robert C, Abergel C, Renesto P, Ogata H, *et al.* (2004). The 1.2-megabase genome sequence of Mimivirus. Science 306:1344–50.
Raoult D, Forterre P. (2008). Redefining viruses: lessons from Mimivirus. Nature reviews. Microbiology 6:315–9.
Rodriguez-Valera F, Martin-Cuadrado A-B, Rodriguez-Brito B, Pasić L, Thingstad TF, Rohwer F, *et al.* (2009). Explaining microbial population genomics through phage predation. Nature Reviews Microbiology 7:828–836.
Rohwer F, Segall A, Steward G, Seguritan V, Breitbart M, Wolven F, *et al.* (2000). The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. Limnology and Oceanography 45:408–418.
Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S, *et al.* (2012). Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. PLoS One 7:e33641.
Sharon I, Alperovitch A, Rohwer F, Haynes M, Glaser F, Atamna-Ismaeel N, *et al.* (2009). Photosystem I gene cassettes are present in marine virus genomes. Nature 461:258–62.
Solioz M, Yen HC, Marris B. (1975). Release and uptake of gene transfer agent by Rhodopseudomonas capsulata. Journal of bacteriology 123:651–7.
Sullivan MB, Coleman ML, Weigele P, Rohwer F, Chisholm SW. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. PLoS biology 3:e144.
Sullivan MB, Huang KH, Ignacio-Espinoza JC, Berlin AM, Kelly L, Weigele PR, *et al.* (2010). Genomic analysis of oceanic cyanobacterial myoviruses compared with T4-like myoviruses from diverse hosts and environments. Environmental microbiology 12:3035–56.
Suttle CA. (2007). Marine viruses--major players in the global ecosystem. Nature Reviews Microbiology 5:801–812.
Thompson LR, Zeng Q, Kelly L, Huang KH, Singer AU, Stubbe J, *et al.* (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. Proceedings of the National Academy of Sciences of the United States of America 108:E757–64.
Vega Thurber R, Haynes M, Breitbart M, Wegley L, Rohwer F. (2009). Laboratory procedures to generate viral metagenomes. Nature protocols 4:470–483.
Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C, *et al.* (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. Proceedings of the National Academy of Sciences of the United States of America 105:18413–8.
Villion M, Chopin M-C, Deveau H, Ehrlich SD, Moineau S, Chopin A. (2009). P087, a lactococcal phage with a morphogenesis module similar to an Enterococcus faecalis prophage. Virology 388:49–56.
Wahba AJ, Miller MJ, Alain N, Landers TA, Carmichael GG, Weber K, *et al.* (1974). Subunit I of Qβ Replicase and 30S Ribosomal Protein S1 of Escherichia Coli. The journal of biological chemistry 249:3314–3316.
Wommack KE, Bhavsar J, Ravel J. (2008). Metagenomics: read length matters. Applied and environmental microbiology 74:1453–1463.
Wommack KE, Colwell RR. (2000). Virioplankton: viruses in aquatic

ecosystems. Microbiology and Molecular Biology Reviews 64:69–114.
Yen HC, Hu NT, Marrs BL. (1979). Characterization of the gene transfer agent made by an overproducer mutant of Rhodopseudomonas capsulata. Journal of molecular biology 131:157–168.

Impact des méthodes de préparation et des contaminations potentielles sur les résultats d'analyses comparatives de viromes

Au final, l'évaluation de la contamination des viromes par des séquences issues de génomes cellulaires a permis de montrer que de tels cas étaient rares, et que les jeux de données de métagénomique virale comportaient majoritairement des séquences issues de

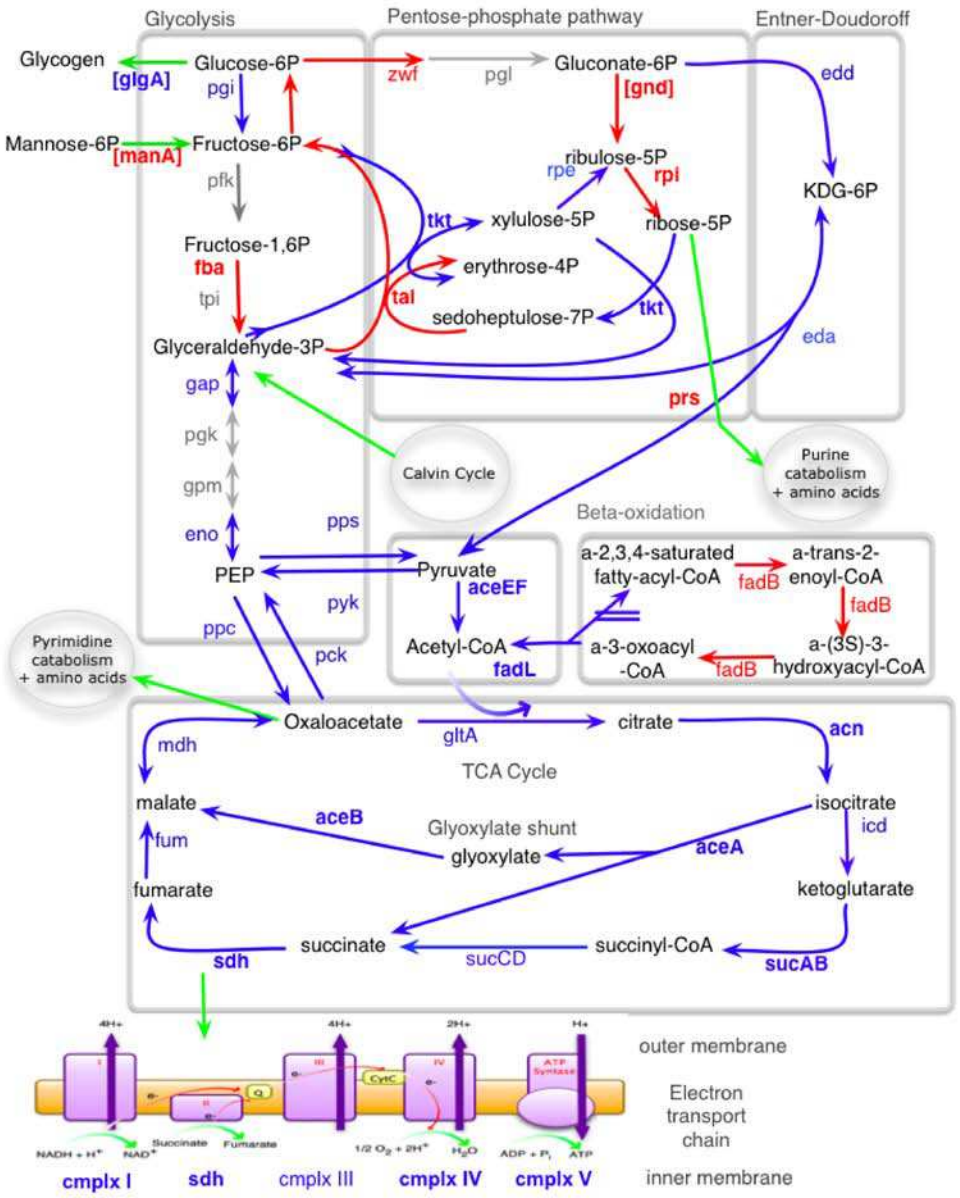


Figure II.2 : Vue d'ensemble des cycles associés au métabolisme du carbone et des enzymes retrouvés au sein des viromes océaniques de l'Océan Pacifique. *Le nom de chaque enzyme est (i) coloré en fonction de la zone dans laquelle l'enzyme est majoritairement retrouvée (rouge : photique, bleu : aphotique, gris : non détecté), (ii) entre crochets si présente dans l'ensemble des viromes POV, et (iii) en gras si retrouvé dans plus de 300 séquences. Des flots de carbones putatifs associés aux zones photiques et aphotiques sont indiqués par les flèches colorés, les entrées et sorties étant indiquées en vert (Hurwitz, Hallam & Sullivan, communication personnelle).*

génomes viraux. Toutefois, la présence parfois importante de séquences issues d'organismes cellulaires au sein de certains viromes a fortement influencé les analyses comparatives de fonctions, puisque l'omission de ces quelques viromes permet de mettre en avant des différences très importantes entre le potentiel fonctionnel des communautés virales et microbiennes, à l'inverse des résultats obtenus en considérant l'ensemble des jeux de données. La gamme de fonctions codées par les génomes viraux a tout de même pu être étendue, avec la détection au sein de viromes non contaminés de plusieurs gènes impliqués dans les métabolismes du carbone, des acide aminés, mais aussi associés aux protéines ribosomales. Ainsi, les génomes viraux ne possèdent sans doute pas une versatilité aussi importante que ceux des micro-organismes, mais le nombre de fonctions détectées avec un échantillonnage encore limité semble indiquer qu'il existe un ensemble de gènes de métabolisme au sein des génomes des virus de l'environnement plus large que ce que décrivent les génomes complets disponibles actuellement.

L'étude fonctionnelle des viromes océaniques POV (Virus de l'Océan Pacifique) a par ailleurs confirmé les résultats obtenus lors de cette analyse de viromes non contaminés (Hurwitz, Hallam & Sullivan, communication personnelle). Plus spécifiquement, cette analyse disposant d'une profondeur de séquençage inédite pour des viromes marins a notamment relevé la présence importante de gènes associés au métabolisme du carbone (Figure II.2). La collection de gènes détectés au sein des viromes semble confirmer que les phages modifient ce cycle cellulaire primordial, avec visiblement une différence entre les enzymes retrouvées en zone photique et aphotique.

Cette méta-analyse de viromes met également en avant les problématiques liées à la préparation des échantillons en vue d'étude de métagénomique virale, et notamment les difficultés rencontrées lors de la purification des capsides. Si certaines contaminations sont clairement irrémédiables, comme par exemple celles associées aux agents de transfert de gènes bactériens (GTA) qui seront systématiquement précipités avec les capsides virales, certains types d'échantillons semblent particulièrement difficiles à traiter par les protocoles actuels. Ainsi, les prélèvements effectués au sein d'organismes eucaryotes, comme le mucus de poumons humains (Willner *et al.*, 2009a), l'intestin de poisson (Dinsdale *et al.*, 2008), ou les fragments de coraux (Vega Thurber, 2009) sont tous composés en partie de séquences d'origine microbienne.

Le cas des coraux en particulier est assez caractéristique des limites et difficultés liées à l'analyse de viromes. Les récifs coraliens sont depuis plusieurs années maintenant soumis à des stress importants, qui entraînent un ensemble de pathologies, notamment le blanchiment de ces coraux (ou "*coral bleaching*", (Brown, 1997; Douglas, 2003)). Certains auteurs ont émis l'hypothèse d'un impact des virus sur ces coraux et leurs organismes symbiotes, et des

approches métagénomiques ont ainsi été appliquées pour caractériser la flore virale associée aux coraux (Marhaver *et al.*, 2008; Vega Thurber *et al.*, 2009b; Correa *et al.*, 2013). Toutefois, ces analyses étant réalisées souvent sur différentes espèces de coraux, avec des protocoles de préparation des échantillons généralement différents entre les études et malheureusement encore imparfaits, les liens éventuels entre les communautés virales observées et les atteintes des coraux restent encore à éclaircir. Notamment, une phase de standardisation des méthodes et analyses est nécessaire avant de pouvoir réellement aborder les questions de l'impact des virus sur les coraux (Annexe A.3).

Au final, les études de viromes ont indéniablement pris une grande part dans l'élaboration d'une image de plus en plus précise de la distribution des virus dans les différents milieux et de leur capacité à transférer et sélectionner des gènes d'intérêt, y compris impliqués dans les cycles métaboliques cellulaires. S'il existe un certain nombre de biais ou de problèmes méthodologiques auxquels il est indispensable d'être attentif, notamment lors d'analyses générales de type analyse fonctionnelle des communautés, les viromes se sont révélés être des outils puissants pour d'appréhender les communautés virales issues de différents types d'échantillons dans leur globalité.

Chapitre III – Facteurs structurants des communautés virales aquatiques

Les facteurs expliquant la distribution des micro-organismes sont à l'heure actuelle sujet à débat. Un courant de pensée principal s'est développé autour de l'ubiquité de ces micro-organismes, synthétisé dès 1934 par Baas Becking dans la formule “tout est partout ; mais l'environnement sélectionne”, qui pourrait expliquer les observations répétées d'espèces microbiennes proches au sein d'environnements aux conditions similaires mais géographiquement très éloignés (Fenchel & Finlay, 2004; de Wit & Bouvier, 2006; Pommier *et al.*, 2012). Toutefois, certains micro-organismes semblent avoir une dispersion limitée par des barrières physiques, et plusieurs études ont ainsi contesté l'ubiquité des micro-organismes (Whitaker, 2006; Pérez-del-Olmo *et al.*, 2009). Parmi les facteurs expliquant les différences entre milieux aquatiques au-delà de la distance géographique, la salinité a plusieurs fois été identifiée comme l'un des facteurs structurant les communautés biologiques, notamment microbiennes (Williams, 1998; Logares *et al.*, 2009, 2012; Bråte *et al.*, 2010)

Au niveau des communautés virales, la distribution des virus au niveau mondial et les paramètres influençant cette distribution ne sont pas encore caractérisés (Vega Thurber, 2009). Il est toutefois raisonnable de penser que de par leur taille et leur nombre, les particules virales sont susceptibles de se disperser sur une aire très large. Une dispersion large des virus permet également d'expliquer l'observation de richesses locales extrêmement élevées tout en faisant l'hypothèse d'une richesse globale raisonnable (Rohwer, 2003) (Breitbart & Rohwer, 2005). Différentes études ciblant des groupes viraux spécifiques ont ainsi été réalisées *via* des approches de PCR, et ont pu confirmer l'existence de séquences quasiment identiques entre des échantillons physiquement éloignés (Breitbart *et al.*, 2004b). Le suivi temporel de communautés virales de sources d'eau chaude physiquement séparées a par ailleurs révélé que la forte diversité locale était maintenue par des migrations entre les différents environnements plutôt que par une accumulation de mutations (Snyder *et al.*, 2007). Toutefois, les résultats obtenus jusqu'à maintenant restent trop partiels (spécifiques d'un groupe viral ou d'un type d'environnement) pour pouvoir généraliser ces hypothèses et déduire des tendances propres aux communautés virales dans leur globalité.

L'analyse de viromes devrait permettre de mieux appréhender cette diversité génétique des communautés virales aquatiques, dans un premier temps en permettant de mieux décrire et caractériser ces communautés dans différents milieux, puis *via* une comparaison de ces viromes afin de révéler d'éventuels paramètres expliquant la distribution des différents virus.

Analyse métagénomique de communauté virales lacustres

S'inscrivant dans le contexte de la description des génomes viraux de l'environnement, le projet METAVIR (programme EC2CO, responsable : Didier Debroas) s'est attaché à l'étude par approche métagénomique des communautés virales de deux lacs écologiquement contrastés : les lacs Pavin et Bourget. Le Lac Pavin, situé dans le massif des Monts Dore (Puy-de-Dôme) à une altitude de 1 197 m, est un lac de cratère alimenté par différentes sources superficielles auxquelles viennent s'ajouter l'apport de sources sous-lacustres. L'importance de sa profondeur (profondeur maximale de 95 m) par rapport à sa surface (0,44 km2) et les caractéristiques topologiques de son bassin versant en font un lac très protégé des vents, et le lac Pavin est ainsi considéré comme méromictique : les eaux de surface sont brassées deux fois par an, tandis que la partie profonde du lac (à partir de 60 m environ) n'est jamais mélangée. Diverses études ont amené à considérer le lac Pavin, peu touché par les activités humaines, comme oligo-mésotrophe (soit relativement pauvre en nutriments ; (Boucher *et al.*, 2006)). Le lac du Bourget, situé au pied des Alpes nord-savoyardes à une altitude de 231 m, et considéré comme le plus grand lac naturel entièrement français (45 km2 et 3,6 milliards de m3 d'eau), a à l'inverse subi de forts impacts anthropiques. Atteint par l'eutrophisation (présence importante voire excessive de nutriments) dans les années 1970, un important programme de restauration de ses eaux a été mis en place dans les années 1980. Désormais, son statut trophique est considéré comme mésotrophe.

Au-delà de leurs spécificités écologiques, ces lacs présentaient l'intérêt d'avoir déjà fait l'objet d'études des communautés virales, que ce soit au niveau de dynamiques temporelles (Bettarel *et al.*, 2003b; Personnic *et al.*, 2009a), de mesures d'activité des virus (Bettarel *et al.*, 2004; Colombet *et al.*, 2006; Personnic *et al.*, 2009b), ou encore d'estimation de la mortalité liée à cette activité virale (Bettarel *et al.*, 2003a; Jacquet *et al.*, 2005). Une analyse ciblée des cyanophages de type T4 avait également été réalisée sur le lac du Bourget (Dorigo *et al.*, 2006), et avait révélé l'existence de différents clades lacustres au sein de ces phages, dont certains très proches de clades marins. En revanche, il n'existait pas au moment de l'étude de virome décrivant des communautés virales ADN de lacs tempérés, les échantillons les plus proches étant issus de piscicultures (Rodriguez-Brito *et al.*, 2010), d'un lac en Antarctique (López-Bueno *et al.*, 2009), et de la communauté de virus à ARN d'un lac Nord-américain (Djikeng *et al.*, 2009). Dans ce contexte, l'analyse métagénomique de communautés virales des lacs Pavin et Bourget devait apporter un regard nouveau sur l'origine, la distribution et la spécificité des virus de lacs tempérés.

Article IV

Assessing the diversity and specificity of two freshwater viral communities through metagenomics.

Simon Roux[1,2], Francois Enault[1,2*], Agnes Robin[1,2], Viviane Ravet[1,2], Sebastien Personnic[1,2], Sebastien Theil[1,2], Jonathan Colombet[1,2], Telesphore Sime-Ngando[1,2], Didier Debroas[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand
[2]CNRS, UMR 6023, LMGE, F-63177 Aubière

Publié le 14 mars 2012 dans **Plos One** (7 (3) : e33641)

Matériel supplémentaire : Annexe A.5

OPEN ACCESS Freely available online

PLoS one

Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics

Simon Roux[1,2], Francois Enault[1,2]*, Agnès Robin[1,2], Viviane Ravet[1,2], Sébastien Personnic[1,2], Sébastien Theil[1,2], Jonathan Colombet[1,2], Télesphore Sime-Ngando[1,2], Didier Debroas[1,2]

1 Laboratoire "Microorganismes: Génome et Environnement", Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France, **2** CNRS, UMR 6023, LMGE, Aubière, France

Abstract

Transitions between saline and fresh waters have been shown to be infrequent for microorganisms. Based on host-specific interactions, the presence of specific clades among hosts suggests the existence of freshwater-specific viral clades. Yet, little is known about the composition and diversity of the temperate freshwater viral communities, and even if freshwater lakes and marine waters harbor distinct clades for particular viral sub-families, this distinction remains to be demonstrated on a community scale. To help identify the characteristics and potential specificities of freshwater viral communities, such communities from two lakes differing by their ecological parameters were studied through metagenomics. Both the cluster richness and the species richness of the Lake Bourget virome were significantly higher that those of the Lake Pavin, highlighting a trend similar to the one observed for microorganisms (i.e. the specie richness observed in mesotrophic lakes is greater than the one observed in oligotrophic lakes). Using 29 previously published viromes, the cluster richness was shown to vary between different environment types and appeared significantly higher in marine ecosystems than in other biomes. Furthermore, significant genetic similarity between viral communities of related environments was highlighted as freshwater, marine and hypersaline environments were separated from each other despite the vast geographical distances between sample locations within each of these biomes. An automated phylogeny procedure was then applied to marker genes of the major families of single-stranded (*Microviridae*, *Circoviridae*, *Nanoviridae*) and double-stranded (*Caudovirales*) DNA viruses. These phylogenetic analyses all spotlighted a very broad diversity and previously unknown clades undetectable by PCR analysis, clades that gathered sequences from the two lakes. Thus, the two freshwater viromes appear closely related, despite the significant ecological differences between the two lakes. Furthermore, freshwater viral communities appear genetically distinct from other aquatic ecosystems, demonstrating the specificity of freshwater viruses at a community scale for the first time.

Citation: Roux S, Enault F, Robin A, Ravet V, Personnic S, et al. (2012) Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. PLoS ONE 7(3): e33641. doi:10.1371/journal.pone.0033641

Editor: Darren P. Martin, Institute of Infectious Disease and Molecular Medicine, South Africa

Received December 23, 2011; **Accepted** February 14, 2012; **Published** March 14, 2012

Funding: This study was funded by an EC2CO CNRS program grant (Metavir). SR was supported by a PhD grant from the French defense procurement agency (DGA, Direction Générale de l'Armement). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: francois.enault@univ-bpclermont.fr

Introduction

Despite the large population sizes of microbes, their high reproductive rates and the potential for long-distance passive dispersal, an increasing amount of studies are showing that the transitions between marine and fresh waters are infrequent [1]. Indeed, marine and freshwater microbes are usually not closely related, often grouping into distinct marine and freshwater clades among bacteria [2] or eukaryotes [3]. Based on host-specific interactions, the presence of specific clades among hosts suggests the existence of freshwater-specific viral clades. Despite the paucity of molecular data from freshwater viruses, recent studies comparing freshwater and marine viruses have concluded on the existence of distinct clades [4,5]. Nevertheless, these PCR-mediated analyses are restricted to chosen viral groups as no gene is universally conserved among viruses. In addition, part of the existing diversity of the viral groups studied is missed as PCR primers are based on previously known sequences described in public databases. Thus, both the diversity of freshwater viral communities and its distinction with marine viruses still need to be demonstrated on a community scale by studying not one but all the major viral families.

Viral metagenomics is a methodology capable of providing an exhaustive view of viral diversity [6], and it has so far revealed an important unknown diversity and an unexpected richness of viral communities [7]. Virome studies on freshwater environments were conducted on aquaculture facilities [8], and a polar lake in Antarctica [9], but never on temperate freshwater lakes. The viral diversity retrieved in these analyses was contrasted : aquaculture facilities were mainly composed of bacteriophages (*Myoviridae* and *Podoviridae*), whereas eukaryotic viruses, including phycodnaviruses and single-stranded DNA viruses accounted for a large proportion of the Antarctic viral communities. Thus, a fine-grained and

exhaustive description of viral diversity in temperate freshwater lakes is needed to improve our knowledge about these communities and to offer the opportunity of identifying potential viral populations specific of those environments.

To assist in these goals, we performed a characterization of freshwater viromes from two french lakes: the lake Pavin and the lake Bourget which exhibit different trophic status, morphological and hydrological features (Table S1). Because species compositions and abundances of potential hosts have been shown to vary with lake trophic status, depth, watershed or size [3,10], suggesting possible similar variations for viral communities, these two lakes are expected to be complementary systems for studies on freshwater viromes.

The two viromes were analyzed according to the following procedure : (i) the characteristics and richness indices of the two viral communities were determined, (ii) freshwater viromes were cross-compared to a set of previously published viromes in terms of sequence similarity and richness, (iii) the composition of the two communities were determined and phylogenetic analyses of the major families were computed in order to accurately describe the genetic diversity in these families.

Results

Overview of the two freshwater viromes

After 454 pyrosequencing and data filtering, viromes of 593,084 and 649,290 reads with an average length of 420 bp were available for Lake Bourget and Lake Pavin, respectively. The proportion of reads similar to protein sequences of the non-redundant NCBI database (NR) were 26.4% and 14.3% for Lakes Bourget and Pavin, respectively (Figure 1A). These proportions of "known" reads (reads with a BLAST hit against NR) are among the highest compared to published viromes (range 1%–28% with an average of 6,3% for aquatic environments [9,11]). Yet, as read length influences these proportions of "known" reads [12] and as our reads are 400 bp versus 100 to 250 bp in previous studies, a direct comparison of BLAST hit ratios is questionable, so the "known" fractions were also determined using reads randomly reduced to 100 bp. Using shorter reads, the "known" fractions dropped to 2.2% in Lake Bourget and to 0.7% in Lake Pavin, the one of the Lake Pavin being the lowest among aquatic viromes (Table S2).

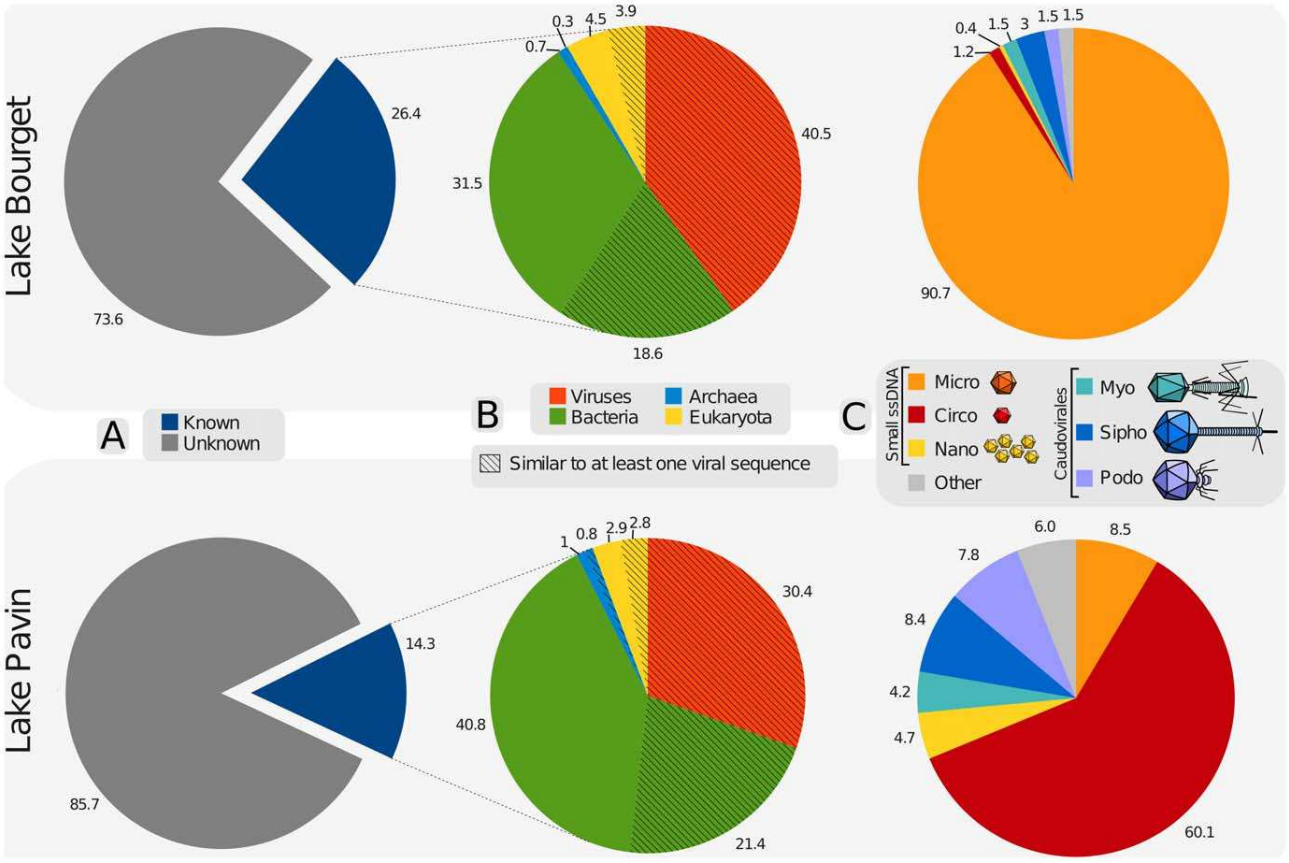


Figure 1. Composition and taxonomic affiliations of Lake Bourget and Lake Pavin virome reads as determined by similarity to known sequences. (A) The percent of "known" virome sequences when compared to the NR protein database. A read was considered "known" if it had a significant similarity in NR (BLASTx using thresholds of 10^{-3} on e-value and 50 on bit score). (B) Breakdown of the "known" sequences into Viruses, *Bacteria*, *Archaea*, or *Eukarya* using similarity results against NR. Hatched parts were reads having a best BLAST hit against a non-viral sequence, but still presenting significant similarity against a complete virus genome sequence of the RefseqVirus database (tBLASTx using thresholds of 10^{-3} on e-value and 50 on bit score) and thus designated as reads "similar to at least one viral sequence". (C) Taxonomic composition at the viral family level of these reads "similar to at least one viral sequence" computed using the GAAS pipeline. The "Other" category pools families which represented less than 1% of the full virome sequences. The number of sequences represented in each chart are as follow : 593,084 ; 156 772 ; 95,905 for charts A, B and C of Lake Bourget virome, and 649,290 ; 92,834 ; 47,345 for the Lake Pavin virome.
doi:10.1371/journal.pone.0033641.g001

Among these "known" fractions, a majority of reads (69.6% for Lake Bourget and 59.5% for Lake Pavin) was most similar to non-viral sequences (Figure 1B), whereas our viromes were not contaminated by bacteria : the absence of 16S rRNA in both samples was checked by PCR amplification and a BLAST search for 16S rRNA sequences, and only a paucity of viromes reads were found to be partly similar to ribosomal proteins (16 reads in Lake Bourget virome, 6 reads in Lake Pavin virome). This result is consistent with previous studies [7] and is thought to be an indication of both the lack of viral gene annotation and the horizontal gene transfers between viral and host genomes.

The taxonomic composition deduced from the bacterial best BLAST hits was not consistent with the previously published data on bacterial communities from Lakes Bourget and Pavin ([10,13] ; Table S3). In the two lakes, members of phylum *Actinobacteria* are considered as dominant in the bacterial fraction, but scarcely retrieved in the viromes. Conversely, a large proportion of virome reads are associated to *Firmicutes* and *Gammaproteobacteria*, two groups considered as rare or even absent in the lakes.

Finally, the gene functions retrieved in the two viromes were unambiguously related to the viral life cycle, with a large proportion of genes related to DNA metabolism, DNA replication, and capsid assembly (Table S4). On the 30 most retrieved PFAM domain, only one (PF05792, Candida agglutinin-like protein) was never found in a viral genome. However, considering that this protein is involved in cell-cell interaction, its presence in viral genomes is rather logical.

Cluster and species richness in the two freshwater viromes

The virome from the Lake Bourget has a higher number of different clusters, as 590,000 randomly chosen reads from each virome were grouped in 272,948 and 113,454 different clusters for Bourget and Pavin, respectively. This higher cluster richness observed in viruses from the Lake Bourget is also revealed by rarefaction curves, as the viral community from Lake Pavin was almost entirely contained in the metagenome, whereas that from Lake Bourget is far from being covered by the virome (Figure S1). The species richness, assessed using PHACCS [14], was estimated to 43,236 and 29,936 different virotypes in Lake Bourget and Lake Pavin, respectively.

Comparison of the richness of 31 viromes. Species richness (number of different virotypes) and cluster richness estimated from the two viromes were compared to the ones of 29 previously published viromes [9,11] from seawater, freshwater, hypersaline and from viral communities associated with different eukaryotes (fish, coral and mosquito) (Table S2). Because numbers of reads of these 31 viromes were different, subsamples of 50,000 randomly-chosen 100-bp reads were generated in order to work with cross-comparable results. Number of virotypes was estimated using PHACCS [14], and plotted for each type of environment (Figure 2A). No significant link could be found between sample origin and virome species richness (one-way ANOVA: p-value = 0.542). The cluster richness of each virome, deduced from the read clustering (Figure 2B), was significantly different between the different environments (one-way ANOVA: p-value = 0.035). Genetic diversity of aquatic viromes (whether marine, freshwater, or hypersaline) was higher than that of viromes associated with eukaryotes, the highest diversity being observed in marine environment.

Similarity-based comparison between viromes. Classically, bacterial metagenomes are compared using the taxonomic composition of their "known" fraction. Such a comparison would be misleading for viromes as their "known" fraction, generally lower than 10% of the reads, is not representative enough. Thus, the 31 viromes were not compared to a reference database but to each other in order to detect potential genetic links between the different viral communities. The hierarchical clustering tree, computed from tBLASTx comparisons, highlighted a separation of the viral communities according to four environment types: eukaryote-associated, high-salinity water (high and medium hypersaline), low-salinity water (seawater and low-hypersaline), and freshwater (Figure 3). Among freshwater viral communities, viromes from Lake Limnopolar were aggregated and distant from the other viromes, highlighting the specificity of these viral communities. Temperate freshwater viromes were also clustered, and split into two sub-groups: a group composed of the viromes from Lake Bourget and Lake Pavin and a group of viromes from aquaculture facilities [8].

Composition of viral communities and subsequent phylogenies on the main identified viral families

General taxonomic composition. To characterize the composition of viral communities, reads similar to a known viral proteins were affiliated using best BLASTx hits against RefSeqVirus. These identified viral fractions represented 15% and 10 of the reads in Lake Bourget and Lake Pavin viromes, respectively (Figure 1C). The same viral families (Microviridae, Circoviridae, Nanoviridae, Myoviridae, Siphoviridae and Podoviridae) were retrieved in both lakes but in different relative proportions (Figure 1C). As previously reported in other ecosystems [9,15], a high proportion of single-stranded DNA (ssDNA) viruses (Micro-, Circo- and Nanoviridae) virus was recorded.

Phylogenetic analysis of the main viral families. Using the procedure described by Roux et al. [16], phylogenetic analyses were performed on the major viral families of the two viromes using different marker genes. Virome reads homologous to each marker were assembled into contigs long enough to build informative phylogenetic trees for the targeted viral groups.

Phylogenetic analysis of small ssDNA viruses : Microviridae family. *Microviridae* form a group of ssDNA bacteriophages with a small capsid (30 nm), split into two sub-families: *Enterobacteria* phages, and *Gokushovirinae* [17], containing phages of *Chlamydiae*, *Bdellovibrio* and *Spiroplasma*. A set of 892 reads from Pavin and Bourget viromes, similar to the capsid protein VP1, was used to perform phylogenetic analyses. Metagenomic sequences were mainly affiliated to *Gokushovirinae* (78%) forming a clade related to *Chlamydiae* phages but distinct from references and with a high internal diversity, that we named "*Gokushovirinae* : Freshwater clade" (Figure 4). The other metagenomic sequences are distinct from both the *Enterobacteria* phages and the *Gokushovirinae* sub-families (22% of the reads).

Phylogenetic analysis of small ssDNA viruses: Circo- and Nanoviridae families. The replication protein Rep is conserved in different families of small ssDNA eukaryotic viruses (*Circoviridae*, *Nanoviridae*, Satellites viruses, *Chaetoceros* viruses and *Geminiviridae*) and in several recently-sequenced aquatic viruses described as "Circo-like" [18]. 223 assembled virome sequences were included in different part of the multiple alignment of the reference sequences of the Rep protein. Even if no close relationship was evident, sequences from Lake Bourget and Lake Pavin exhibited more similarities with Circo-like sequences from aquatic environments than with other ssDNA viruses. The length of branches in the resulting trees indicates high genetic distances between the different environmental sequences (Figure 5). These results shed light on a hitherto undescribed diversity for this viral family.

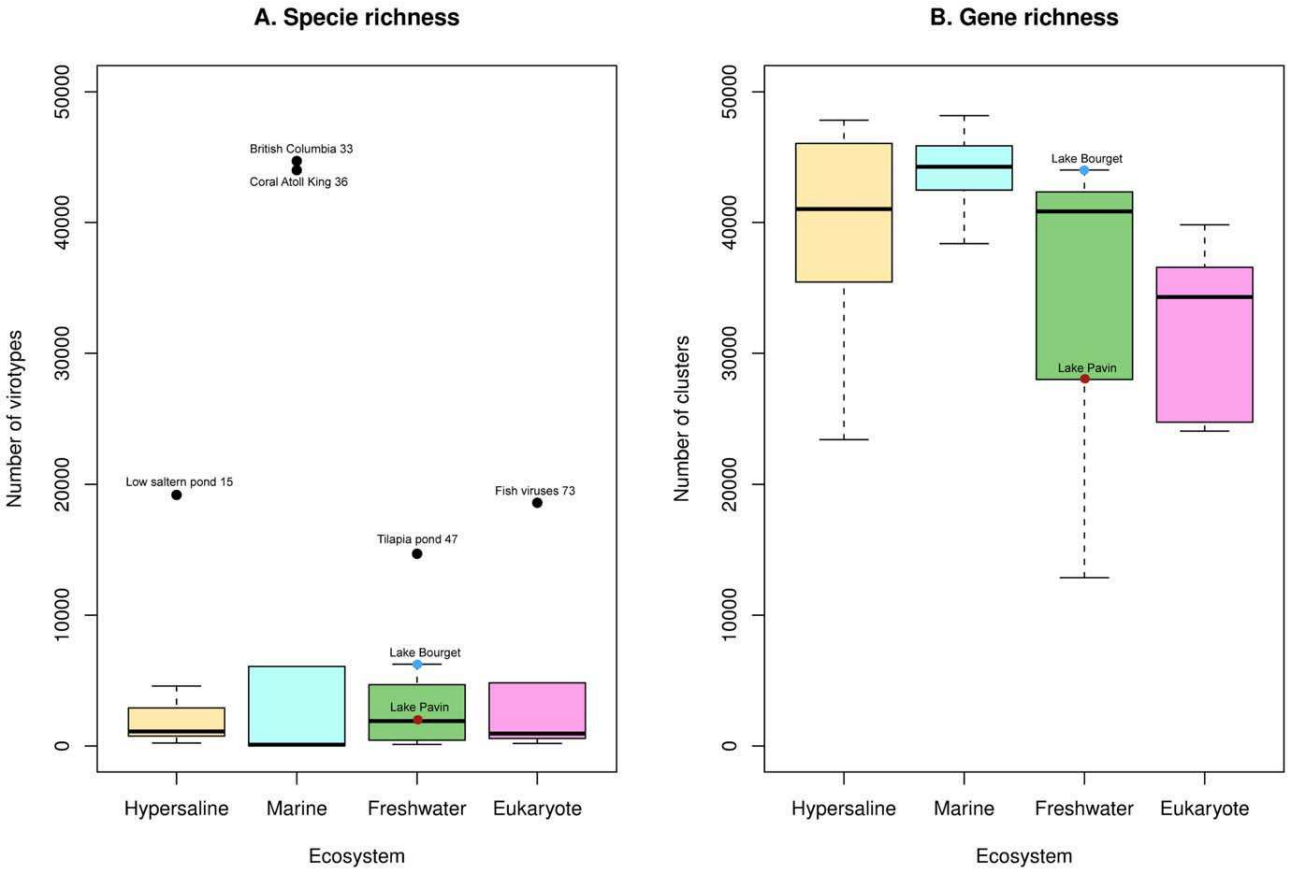


Figure 2. Boxplots of the estimation of species and cluster richness for viromes from different origins. Species richness (A) was estimated with PHACCS for virome subsamples (50,000 sequences, 100 bp). Cluster richness (B) was deduced from the number of different clusters formed from 50,000 input sequences of 100 bp. Viromes associated with extreme points are indicated on the plot, as well as the viromes from Lakes Bourget and Pavin.
doi:10.1371/journal.pone.0033641.g002

Phylogenetic analyses of dsDNA viruses: the Caudovirales. Caudovirales form an order of dsDNA viruses better known as "tailed bacteriophages" whose diversity can be assessed with a gene coding for the large subunit of the terminase (TerL). The 185 virome sequences used in phylogenetic analyses were widely distributed among the *Myo-*, *Sipho-* and *Podoviridae* and were phylogenetically distant to known reference sequences, hightlighting an important uncharacterized diversity for Caudoviruses in freshwater environments (Figure 6).

A significant number of these sequences (20%) were related to T4-like phages (Figure 6) within the *Myoviridae* family. The diversity of this group has been previously explored using GP23 and G20 markers [19,20] leading to the identification of different sub-groups: "Near-T4", "T4-like cyanophages" and "Far-T4", a group composed of only one sequenced genome (*Rhodothermus* phage RM378) and identified in marine waters by PCR approach on the GP23 gene [20]. According to G20-based phylogenetic analyses (Figure S2 and Figure S3), 11% of the 190 virome reads from Bourget and Pavin were affiliated to "T4-Like cyanophages" and 89% of these reads formed a new group including *Rhodothermus* phage RM378. Similar proportions were obtained with the GP23 marker, with 16% of the 251 reads affiliated to "T4-Like Cyanophage" and 84% forming a group containing the "Far-T4" group identified by Comeau et al. [20] (Figure S4 and Figure 7). Thus, freshwater virome sequences greatly expand the diversity of the previously identified Far-T4 group.

Discussion

Viruses are the most abundant biological entities in fresh and marine waters, exceeding prokaryotic abundance 10-fold on average. They are important factors for the regulation of microbial community composition [21–23] and affect the cycling of carbon and nutrients [24,25]. Yet, little is known about the composition and diversity of the temperate freshwater viral communities. This study examined such communities from two temperate freshwate lakes differing by their ecological parameters.

Comparison of the two lacustrine communities

The fraction of known reads was higher in the Lake Bourget virome than in the Lake Pavin virome. When reduced to 100-bp reads, the Lake Pavin dataset was the virome with the lowest "known" fraction. These results highlight the lack of knowledge and reference sequences for viruses of these environments, especially for low trophic status waters. Genetic diversity appeared to be high in the Lake Bourget virome, as more than 600,000 reads were still not enough to cover the entire viral community gene pool for this sample. Even if this cluster richness is also high in the Lake Pavin virome, it appeared to be more than twofold lower than the cluster richness of the Bourget. The number of different virotypes (species richness) was also 44% higher in the virome of the Lake Bourget. Species compositions, diversities and abundances of potential hosts have been shown to vary with lake

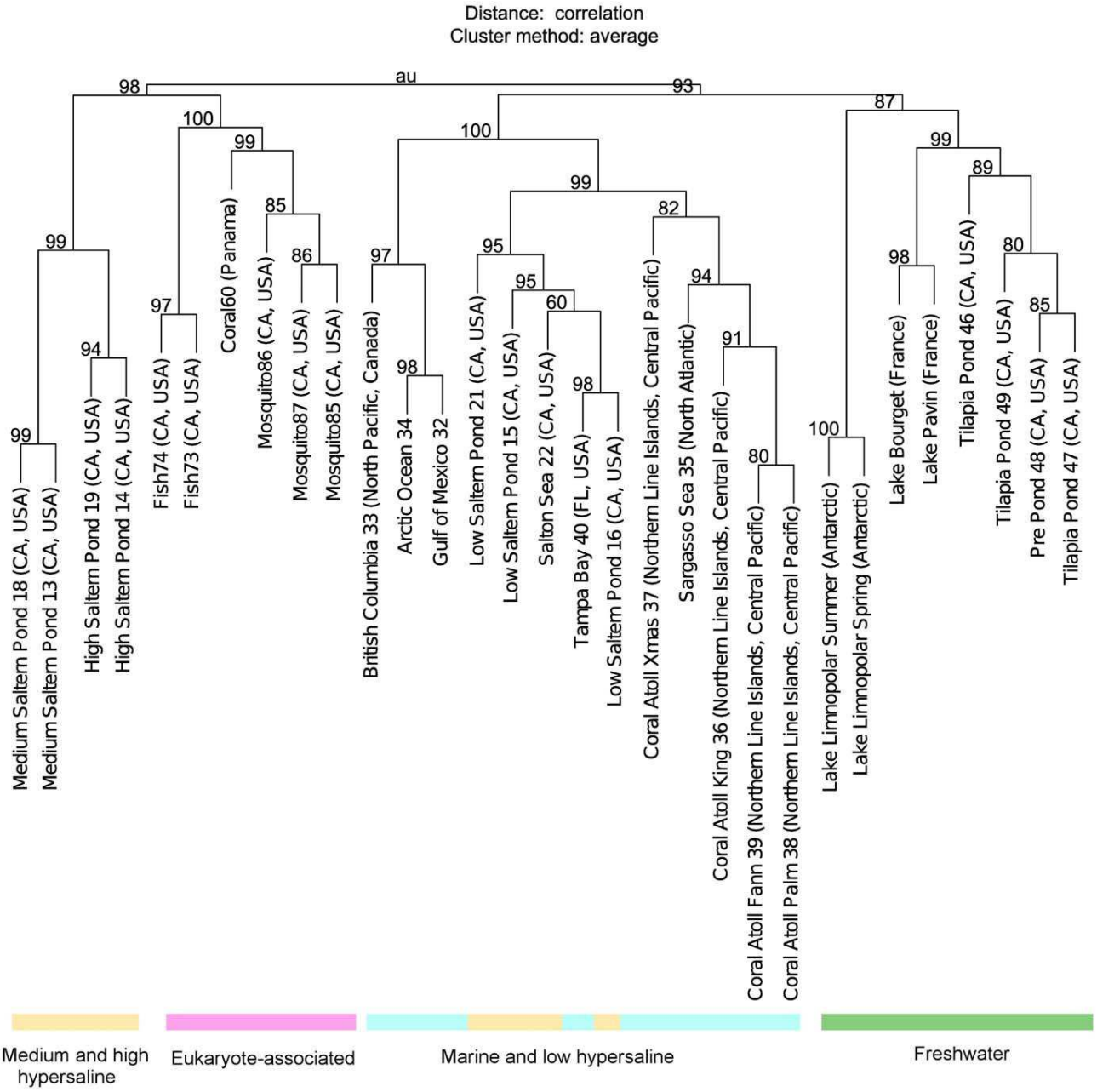


Figure 3. Virome hierarchical clustering tree based on sequence similarity. This tree was computed from tBLASTx comparisons of virome subsamples. Eukaryote-associated viromes were taken from fish, mosquito or coral samples. Hypersaline viromes are split into three categories based on salinity, as indicated in the original study of these viromes (http://www.theseed.org/DinsdaleSupplementalMaterial/).
doi:10.1371/journal.pone.0033641.g003

trophic status, depth, watershed or size [3,10]. Moreover, the richness of bacterial and eukaryotic communities determined in different freshwater ecosystems were lower for low trophic status ecosystems such as Lake Pavin [3]. The analysis of the two viromes, highlighting a similar trend for two freshwater viral communities, corroborates the idea that viral diversity is correlate to the diversity of the organisms seen as potential hosts.

Comparison of aquatic communities

The two freshwater viromes and 29 published viromes were compared through their gene and species richness and through their pairwise sequence similarities. These three analyses are particularly well adapted to viromes as they take into account not only the "known" reads but also the "unknown" reads, that make up the major share of the viral metagenomes in the present state of sequence databases. Species richness highlights the very broad genotypic diversity of viral communities, this richness remaining identical through the different ecosystems studied. Conversely, cluster richness was shown to vary between different environment types and appeared significantly higher in marine ecosystems. These differences could be linked to a difference in genome size (i.e. the same number of genomes would be retrieved in the

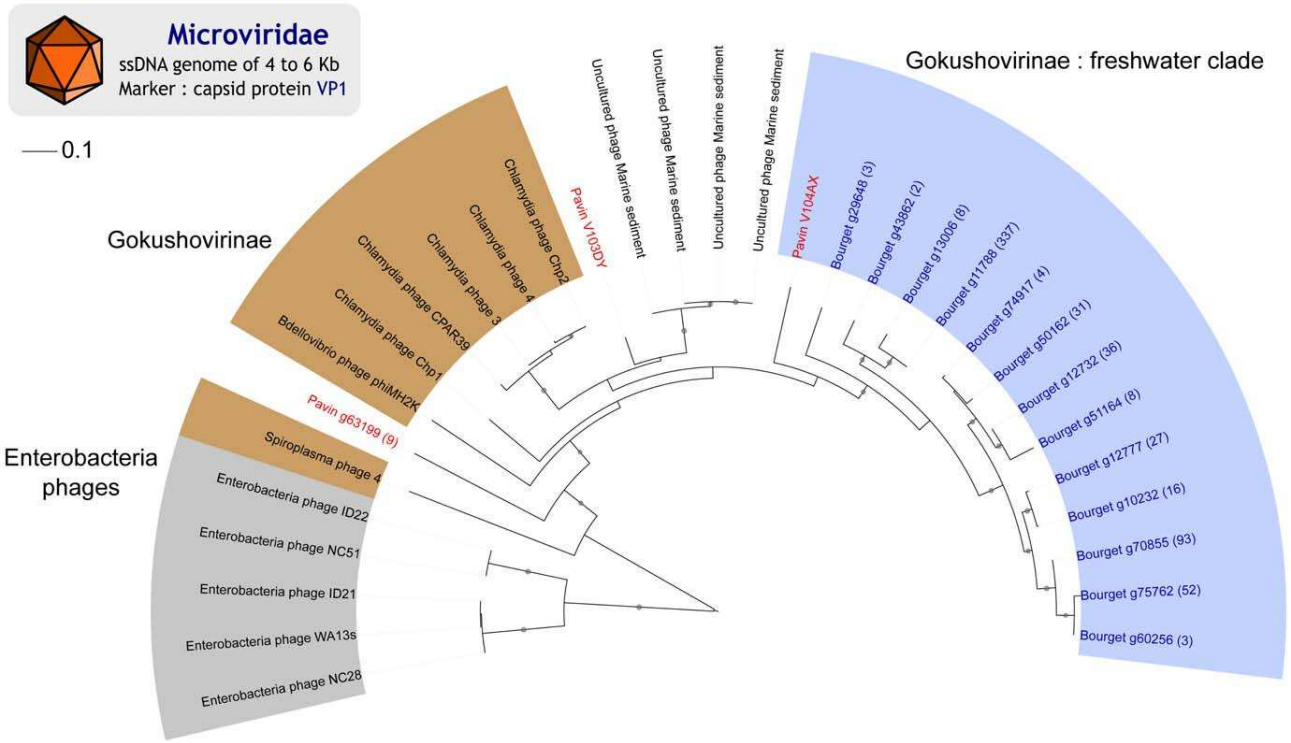


Figure 4. Maximum-likelihood tree for *Microviridae* (VP1). The main reference groups among *Microviridae* were retrieved and indicated on the tree (*Enterobacteria* phages in gray and *Gokushovirinae* in brown). The new group (*Gokushovirinae* : Freshwater clade) is colored in green. Leaf labels corresponding to virome sequences are colored (blue for Lake Bourget and red for Lake Pavin). The number of reads assembled is given in brackets for each contig. Nodes with at least 80% bootstrap support are flagged with black circles.
doi:10.1371/journal.pone.0033641.g004

different ecosystems, but their difference in sizes would lead to a difference in the number of clusters formed), or to a difference in gene and genome richness (in which case the difference in cluster numbers would be linked to a difference in the number of different viral genomes between the ecosystems). Furthermore, viral communities appear to hierarchically cluster according to salinity levels. Specifically, viromes from freshwater environments are clustered, reflecting significant genetic similarity between these viromes, despite the vast geographical distances between sample locations (Antarctica, North America and Europe). Marine environments, as well as hypersaline environments, were also gathered, revealing that viruses are notably distinct in the different aquatic environments. The differences between freshwater and saline aquatic viral communities presented here are consistent with specific studies led on single marker genes [1,5] indicating that freshwater viral communities are specific and distinct from other aquatic viral communities. Moreover, considering that the viromes were not prepared in the same way (e.g. the use of CsCl gradient in some cases, PEG precipitation in others), the result that viromes cluster by salinity level is definitively reflecting a strong pattern of differentiation between aquatic viral communities.

Small ssDNA viruses: an under-estimated group

A strong presence of ssDNA viruses (mostly *Microviridae* and *Circoviridae*) was found in viromes of Lake Bourget and Lake Pavin, both quantitatively and in terms of diversity. These viral groups have been previously identified in different aquatic ecosystems: an analysis of a Sargasso Sea virome described the presence of *Microviridae* in marine environments [26], and *Circoviridae* appeared to dominate Lake Limnopolar viral community [9]. The strong presence of small circular ssDNA viruses is generally thought to be a bias of the genomic amplification [27]. Indeed, small circular DNA fragments, such as genomes of *Circoviridae* and *Microviridae*, were shown to be preferentially amplified by the phi29 DNA amplification process needed to provide enough genetic material for pyrosequencing and used in all the viromes sequenced with NGS [9,28]. However, even allowing for potential quantitative bias, *Microviridae* and *Circoviridae* are undoubtedly present in both Lake Bourget and Lake Pavin.

Identification of a great diversity and of previously unknown clades

Our deep sequencing effort associated with a large read size made it possible to create direct phylogenetic analyses on marker genes from a virome. The main PCR pitfalls, namely the overriding need for a close known sequence in order to design PCR primers, could thus be circumvented. Analyses of the major viral groups found in the two communities all spotlight a very broad diversity and previously unknown virotypes. Indeed, new viral groups were highlighted for each studied family.

Using phylogenetic trees based on the replication protein, two studies recently reported new kinds of *Circoviruses* in aquatic environments that appear to be intermediate between the *Circoviridae*, *Geminiviridae* and *Nanoviridae*, with a very low similarity rate between these new sequences and the references [9,18]. The same pattern emerges from the analysis of the Rep sequences retrieved in Lake Bourget and Lake Pavin. Since hosts of known *Circoviridae* are animals while those of known *Nanoviridae* and *Geminiviridae* are plants, potential hosts for the new viruses are difficult to assess.

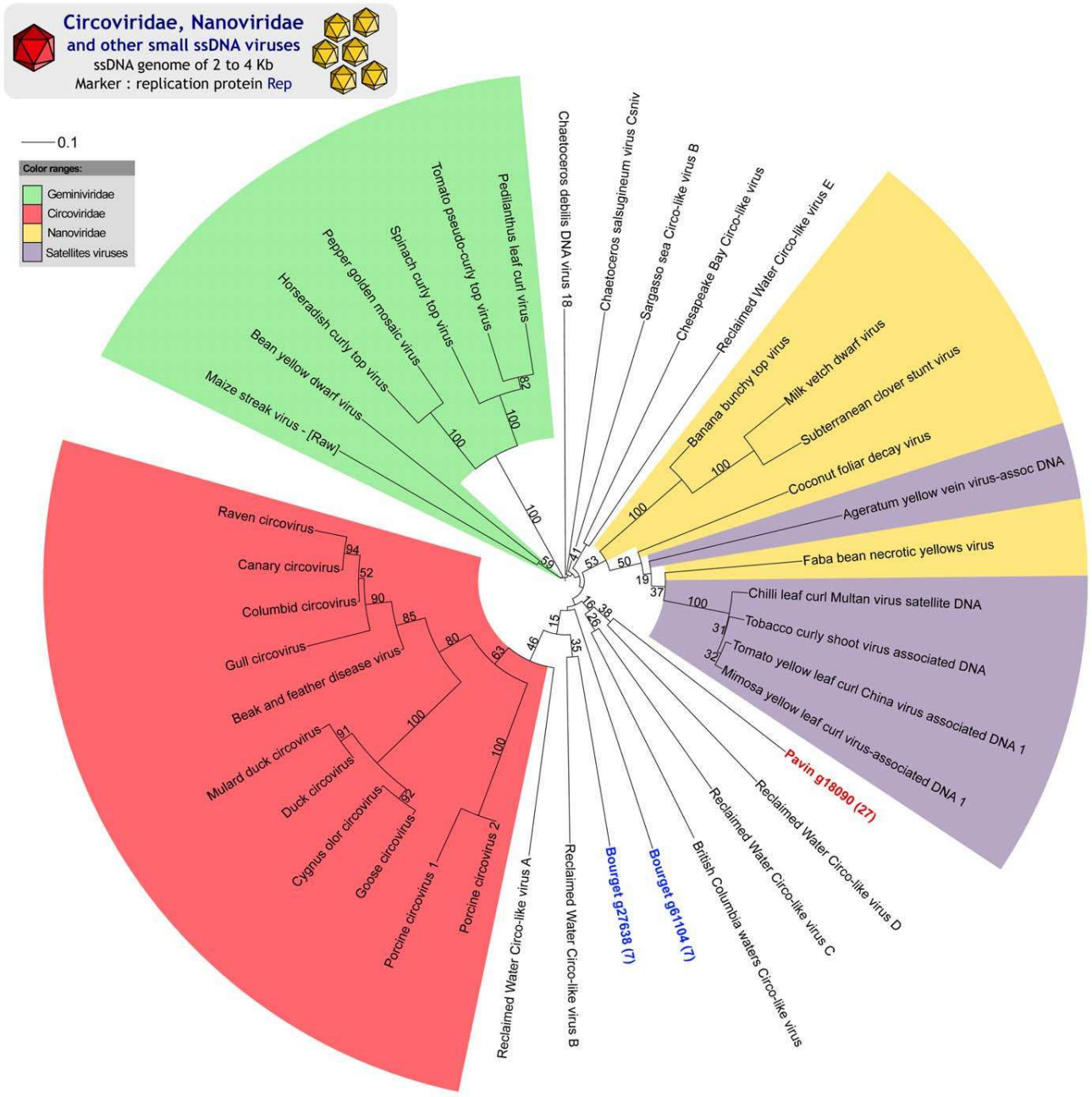


Figure 5. Maximum-likelihood tree for Circo-like viruses (Rep). This tree gathers sequences from four different viral families, i.e. *Circoviridae* in red, *Nanoviridae* in yellow, *Satellites viruses* in blue and *Geminiviridae* in green, sequences from two diatoms viruses (*Chaetoceros* viruses), sequences taken from marine and reclaimed water samples, and sequences from both the lakes studied here. Leaf labels corresponding to virome sequences are colored (blue for Lake Bourget and red for Lake Pavin). Bootstrap support is indicated for each node.
doi:10.1371/journal.pone.0033641.g005

A broad unknown diversity has also been found for *Microviridae* and *Caudovirales*, and putative freshwater clades could be described for these two families. According to our hypothesis, we expected to find freshwater-specific viral clades, due to the presence of typical freshwater clades for microorganism communities [2]. For *Microviridae*, a new group related to the *Chlamydiae* phage and including sequences from Lake Bourget and Lake Pavin is notably distinct from a group of uncultured *Microviridae* sampled from aquatic sedimentary structure [29]. Microphages communities from Lake Bourget and Lake Pavin thus appear to be closely related. These new viruses are unlikely *Chlamydiae* phages, since *Chlamydiae* are rarely detected in lakes and have never been retrieved in previous studies of Lake Bourget [13,30], in the virome of which *Microviridae* were most abundant. Thus, it is likely that these new microphages infect another type of bacteria, or that the *Chlamydiae* phages host range is broader than previously

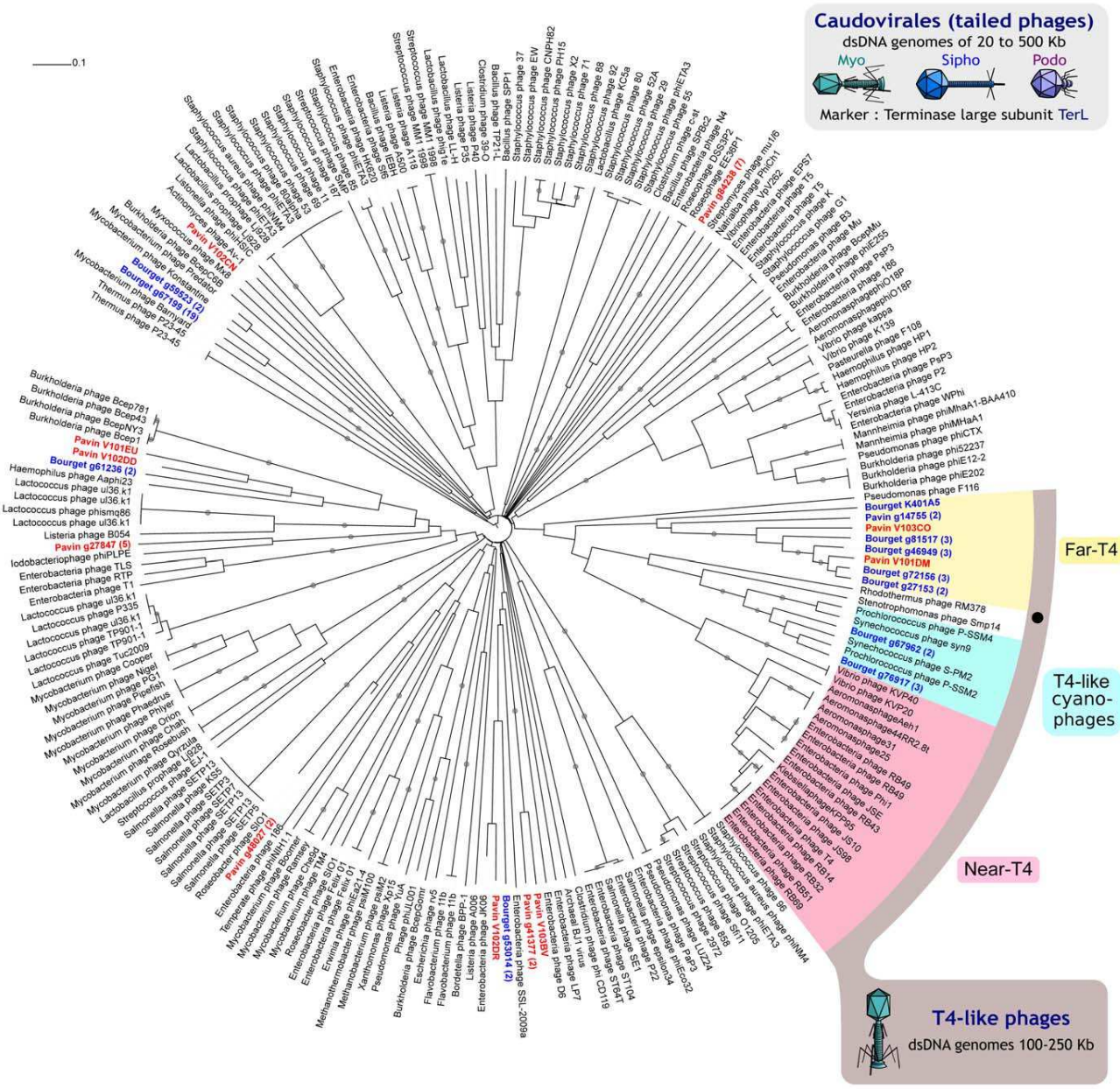


Figure 6. Maximum-likelihood tree for Caudovirales (TerL) : general view (A) and detailed view of the T4-like group (B). The main reference groups of T4-like phages are indicated (near-T4 in red, T4-like cyanophages in blue), and leaf labels corresponding to virome sequences are colored (blue for Lake Bourget and red for Lake Pavin). The Far-T4 group is highlighted in yellow. The number of reads assembled is given in brackets for each contig. Nodes with at least 80% bootstrap support are flagged with black circles. Rhodothermus RM378, the only cultured representative within the Far-T4 clade, is marked with a black dot.
doi:10.1371/journal.pone.0033641.g006

thought. Otherwise, they could actually infect *Chlamydiae*, keeping their numbers below the detection threshold of classic diversity analysis protocols.

Phylogenetic trees drawn from TerL shed light on a high diversity among *Caudovirales*. The major share of the virome sequences is distributed far from references and far from each other, highlighting both the richness of Caudovirales freshwater communities and the absence of closely-related reference sequences.

In addition, some virome sequences appear to form a new clade related to the T4-like viruses, one of the most thoroughly described *Caudovirales* family. The use of degenerate primer sets for GP23 by Comeau *et al.* [20] recently highlighted a group of T4-like phages far from all references, designate as Far-T4 group. In our metagenomic data, phylogenetic analyses on marker genes all indicated the existence of a Far-T4 group, more than 80% of the T4-like phage sequences being affiliated to this group. A comparison of this new viral clade G20 sequences with known PCR primers revealed that these Far-T4 sequences would not be amplified by the primers commonly used to assess the diversity and distribution of T4-like phages in marine environments [31].

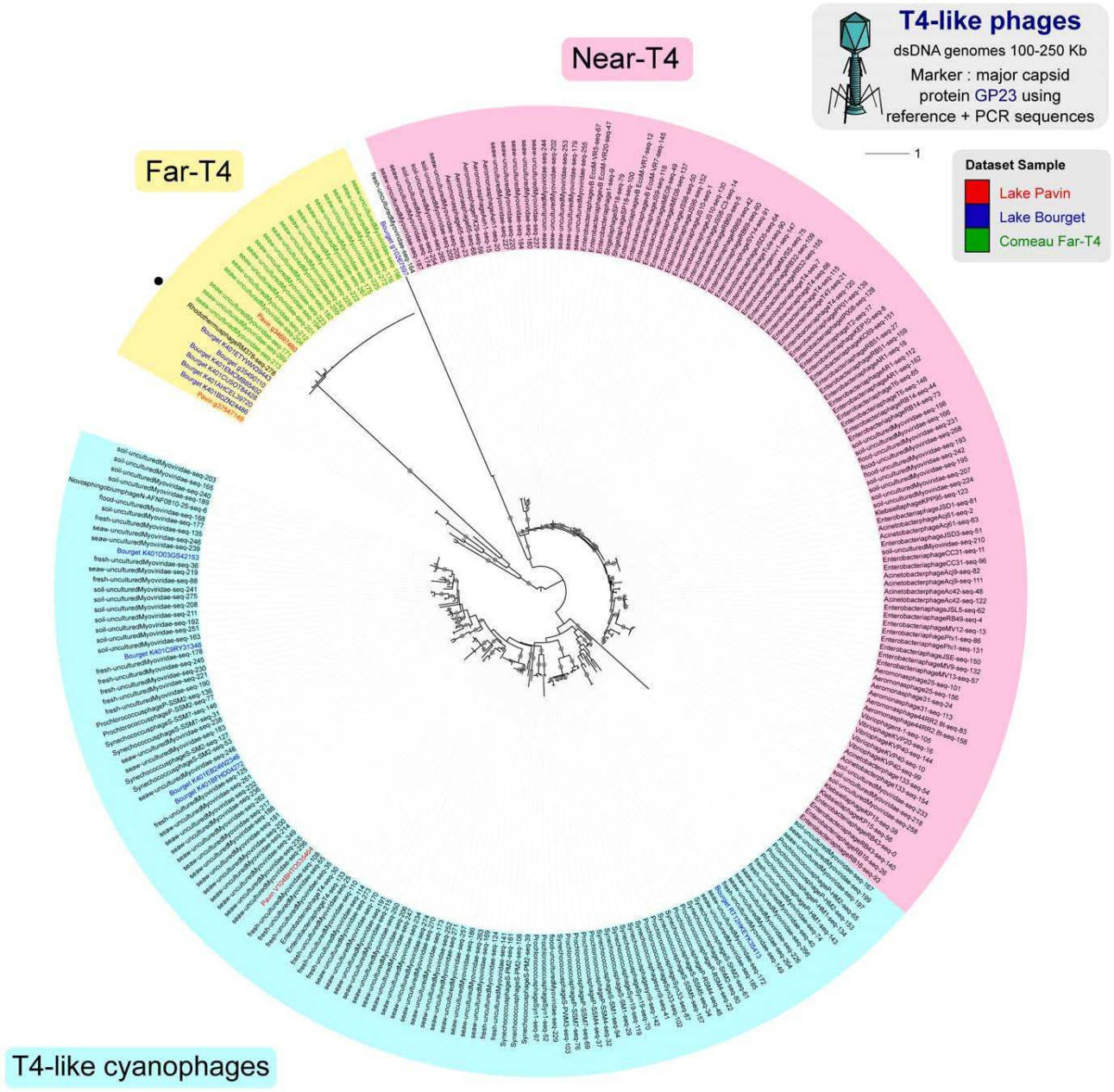


Figure 7. Maximum-likelihood tree for T4-like phage (GP23). The two main reference groups are indicated (near-T4 in red and T4-like cyanophages in blue). The Far-T4 group is highlighted in yellow. Leaves corresponding to virome sequences are colored (red for Lake Pavin and blue for Lake Bourget). Far-T4 sequences described by Comeau *et al.* are highlighted in green. Nodes with at least 80% bootstrap support are flagged with black circles. The sample origin of PCR-obtained sequences is designated on the leaf label (seaw stands for seawater, flood for floodwater, and fresh for freshwater). *Rhodothermus* RM378, the only cultured representative within the Far-T4 clade, is marked with a black dot.
doi:10.1371/journal.pone.0033641.g007

Furthermore, the analysis on GP23 shows that only a part of the diversity of the Far-T4 group is captured by the only primers amplifying DNA sequences from this group [20]. Moreover, this study proves that this Far-T4 group is not specific to marine ecosystems, despite the absence of amplification observed by Comeau *et al.* [20] on different freshwater samples. Thus, this study confirms undoubtedly the existence of a Far-T4 group as such viruses are retrieved using all three markers and indicates that this group is both greatly diversified and quantitatively important in freshwater ecosystems.

Finally, all the observed viral groups contained sequences from both lakes studied here, and phylogenetic trees revealed putative freshwater clades for *Microviridae* and *Caudovirales*, as was expected based on the specificity of microbial clades in lakes. Thus, phylogenetic analysis of major viral groups showed that the two freshwater communities are closely related, despite the significant ecological differences between the two lakes. These two viral communities are thus probably composed of evolutionarily close virotypes, and differ mainly in terms of the relative abundance of the viral species. The specificity of freshwater viruses, allready

known for specific virotypes, is here demonstrated at a community scale and these results call for further studies of this kind on viral communities from a broad spectrum of environments.

Materials and Methods

Sample preparation

Samples were collected from Lake Bourget (45°43′47″N, 5°52′10″E) and Lake Pavin (45°29′41″N, 2°53′13″E) in July and June 2008. Lake Bourget and Lake Pavin are both freshwater lakes but they present significant differences in terms of biophysical parameters : Lake Bourget is ten times bigger than Lake Pavin, is deeper (145 m for Lake Bourget, 92 m for Lake Pavin), and has a greater drainage basin (56,000 ha vs 50 ha). Furthermore, Lake Bourget is a mesotrophic lake, exposed to human activity whereas the oligomesotrophic Lake Pavin is more isolated and located in a former caldera.

Both lakes were sampled by collecting 20 liters of water at a 5 m depth and running serial filtrations (25 µm, 1.2 µm, 0.2 µm). Virus-like particles (VLPs) were concentrated by tangential ultra-filtration (Amicon pump) followed by PEG precipitation [32]. Viral concentrates were then re-filtrated on a 0.2 µm screen, to remove any remaining cellular micro-organisms, then quantified by flow cytometry [33,34]. Final concentration of viral particles was approximately $1.6*10^{10}$ VLPs/ml for both samples, which represented a concentration factor of 1000 from the initial concentration (about 10^7 VLPs/ml for both samples). Viral concentrates were treated with DNAseI (Invitrogen) to remove external DNA fragments.

Encapsidated DNA was freed via thermal shock then purified using a QuiAmp DNA mini kit (Quiagen). To obtain sufficient genomic material for pyrosequencing, DNA amplification was run with a GenomiPhi Kit (GE Healthcare) which produced non-specific amplification through polymerase phi29 (as in [9,28]). Absence of bacterial contamination was checked by flow cytometry [33] before DNA extraction. Potential bacterial contamination was also checked by amplifying the gene coding for 16S rRNA at each purification step (primer set 27f-1492r; [10]). No amplification was found for any of the samples. The two DNA preparations were subjected to a single pyrosequencing run by GATC Biotech (Germany) using a 454 Life Sciences GS-FLX Genome Sequencer.

The datasets generated for the Lake Bourget and the Lake Pavin were composed of 597,675 and 684,224 DNA sequences (i.e. reads) with a mean size of 433 bp and 412 bp, respectively. Replicate software [35] was used to remove exact duplicate reads, which accounted for 6% of the Lake Pavin virome and 1% of the Lake Bourget virome. Both viromes are available through the Short Read Archive under accession number ERP000339, and on the Metavir web-server ([16], http://metavir-meb.univ-bpclermont.fr; project "French lakes").

Public virome dataset

29 viromes available in public databases and composed of more than 50,000 sequences were downloaded for comparison with the viromes of Lakes Bourget and Pavin. 23 originated from aquatic environments and 6 were sampled from eukaryotes (Table S2). All viromes were screened for duplicate reads using Replicate [35]. In order to normalize the sequence size from these different datasets, we sampled 50,000 sequences of 100 bp for each.

Viral communities cluster richness and rarefaction curves

The cluster richness of each virome subsample was assessed by clustering reads. Uclust [36] was used to cluster reads of each virome subsample at 75% identity. This clustering threshold was chosen based on the high divergence observed between viral genes, but similar results were obtained with thresholds of 90% and 98%. For each virome, a sub-sample was iteratively increased by 1.000 randomly selected sequences (without replacement) and clustered at each step using Uclust [36]. The number of clusters formed was plotted as a function of the number of input sequences. The viral cluster richness of the different types of environment were compared by a one-way ANOVA conducted using the R statistical software. For Lake Pavin and Bourget, these clusterings were also computed on whole viromes in order to draw rarefaction curves.

Viral communities species richness

The species richness of each of virome subsamples was computed using the PHACCS tool ([14], http://biome.sdsu.edu/phaccs/), based on the contig spectrum obtained with a sequence assembly at 98% similarity on at least 35 bp, computation of all rank-abundance distribution laws and default parameters. The average genome size was determined using GAAS as in other studies [37]. The viral species richness of the different types of biomes were compared by a one-way ANOVA conducted using the R statistical software.

Similarity-based comparison between viromes

Finally, we ran an *in silico* qualitative comparison between the different viromes based on sequence similarity (tBLASTx comparison) as described in [38] and [16]. Briefly, virome samples (50,000 sequences for each virome) are cross-compared to every other using tBLASTx. A similarity score is deduced, and used to hierarchically cluster viromes using the pvclust package of R software with default parameters [39].

Taxonomic composition of the viromes

After removal of duplicate reads, virome sequences were analyzed without assembly and compared with BLASTx tool [40] against NR, NCBI's non-redundant amino acid sequence database. The best similarity for each virome read was parsed and assigned as "known" if there was a significant similarity to a protein from the NR database (thresholds of 10^{-3} on e-value and 50 on bit score) and else "unknown" (Figure 1A). In a second step, the reads from the "known" group were classified as viral, bacterial, archaeal, or eukaryotic based on their highest similarity (Figure 1B). Reads similar to provirus sequences are often similar to cellular organism sequences. To identify these transferred viral sequences, tBLASTx was used to compare the virome reads against the complete virus genome sequences of the RefseqVirus database. Any read with a significant similarity (thresholds of 10^{-3} on e-value and 50 on bit score) to one of the previous four taxonomic groups of NR that was also similar to a viral sequence of RefSeqVirus was counted as "similar to at least one viral sequence" (Figure 1B). These reads "similar to at least one viral sequence" were affiliated to viral families and the taxonomic composition of each virome was computed using the GAAS pipeline [26], with thresholds of 50% on similarity percent, 20% on query sequence length, and keeping only the top hit for each query (Figure 1C). 16S rDNA absence was checked via a BLASTn of the viromes reads against RDP [41], a 16S ribosomal DNA sequence database. The absence of bacterial contamination was confirmed by the very low number of reads presenting a best BLAST hit against ribosomal proteins (16 for Lake Bourget virome, 6 for Lake Pavin). The bacterial taxonomic composition was based on the reads which best BLAST hit against the NR

database was a bacterial sequence. Functional annotations were deduced from a rpsBLAST against the PFAM database.

Phylogenetic analysis of main viral families

Randomly sequenced metagenomic reads of around 400 bp do not cover the entire marker genes considered (as the markers considered here were between 600 and 1500 bp long). To circumvent this limitation, a custom-designed procedure was developed to automatically generate phylogenetic trees including metagenomic sequences for a marker gene of interest (Figure S5 ; markers used : VP1 for *Microviridae*, Rep for *Circoviridae*, *Nanoviridae* and *Geminiviridae*, TerL, GP23 and G20 for *Caudovirales*).

One marker gene was selected for each ssDNA viral group. For the *Microviridae* family, we used a gene coding for the major viral coat protein VP1, previously used as phylogenetic marker for these viruses [29]. For several other families of small eukaryotic viruses, including *Circoviridae*, *Nanoviridae*, and *Geminiviridae*, we chose the gene coding for a replication protein Rep previously described as a good marker [9,42]. For dsDNA viruses, two types of markers were used : the broad-spectrum marker TerL, previously used to assess diversity among *Caudovirales*, including *Myoviridae*, *Siphoviridae*, and *Podoviridae* [43], and two well-known phylogenetic markers for T4-type phages, the major capsid protein gene GP23 and the capsid assembly protein gene G20 [19,31,44]. The two markers G20 and GP23 were used either with entire gene sequences from completely sequenced phages (Figure S2, Figure S4) or with shorter but more numerous PCR sequences as reference (Figure 7, Figure S3).

Briefly, metagenomic sequences homologous to each marker were retrieved via BLASTx against NR, and assembled using Cap3 [45] (98% identity on 35 bp) in order to have longer sequences at our disposal. Using these stringent assembly parameters makes it possible to group only sequences from the same virotype [14]. These sequences were aligned against a reference alignment, and alignment bounds for each metagenomic sequence were collected and used to define sub-alignements containing several metagenomic sequences. This step is useful to reduce the number of trees to be calculated and makes it possible to generate trees containing several metagenomic sequences. Multiple alignments were automatically curated using Gblocks [46] and the ten longest alignments were selected for each marker. Phylogenetic trees were generated using PhyML [47], with 100 bootstraps replicates. The trees used in the figures were manually edited using iTOL [48].

All these analyses can be viewed on-line on the Metavir web-server ([16], http://metavir-meb.univ-bpclermont.fr). Furthermore, the different Perl scripts designed for these analysis (Virome tBLASTx comparison and automatic tree generation), as well as the multiple alignments and phylogenetic trees generated for this study are available on demand.

Supporting Information

Figure S1 Rarefaction curves based on whole viromes. Each virome was clusterized at 75% identity, and the curve presents the number of different clusters as a function of the number of input sequences.
(PDF)

Figure S2 Maximum-likelihood tree for T4-like phages (G20). The main reference groups are indicated on the tree (near-T4 in red, T4-like cyanophages in blue), and the Far-T4 group is highlighted in yellow. Leaf labels corresponding to virome sequences are colored (red for Lake Pavin and blue for Lake Bourget). The number of reads assembled is given in brackets for each contig. Nodes with at least 80% bootstrap support are flagged with black circles.
(PDF)

Figure S3 Maximum-likelihood tree for T4-like phage (G20). A phylogenetic tree has been drawn for the T4-like phage group, and the two main reference groups are indicated (near-T4 in red and T4-like cyanophages in blue). The Far-T4 group is highlighted in yellow. Leaf labels are colored according to their sample (red for Lake Pavin and blue for Lake Bourget). Nodes with at least 80% bootstrap support are flagged with black circles. The sample origin of PCR-obtained sequences are designated on the leaf label (seaw stands for seawater, flood for floodwater, and fresh for freshwater). Rhodothermus RM378, the only cultured representative within the Far-T4 clade, is marked with a black dot.
(PDF)

Figure S4 Maximum-likelihood tree for T4-like phages (GP23). The main reference groups are indicated on the tree (near-T4 in red, T4-like cyanophages in blue), and leaf labels corresponding to virome sequences are colored (red for Lake Pavin and blue for Lake Bourget). The Far-T4 group is highlighted in yellow. The number of reads assembled is given in brackets for each contig. Nodes with at least 80% bootstrap support are flagged with black circles.
(PDF)

Figure S5 Schematic representation of the phylogenetic tree creation pipeline.
(PDF)

Table S1 Characteristics of the two lakes studied.
(DOC)

Table S2 Main characteristics of viromes included in the different comparisons. Extraction methodology is indicated where available. Peg = polyethylene glycol; CsCl = Cesium Chloride. The BLAST hit ratio is the percentage of reads significantly similar to a protein of the database the non-redundant database (threshold of 10^{-3} on e-value and 50 on scores). As sequence comparison results depend on the length of the sequences, reads were reduced to 100-bp long for all viromes.
(DOC)

Table S3 Bacterial taxonomic composition as deduced from virome reads best BLAST hits, compared with previously published data. These previous data are from a metagenome for Lake Bourget, and from 16SrRNA PCR amplification for Lake Pavin.
(DOC)

Table S4 Main functions retrieved in the viromes. In the first table, the 30 most retrieved PFAM domains in the viromes are listed, with the number of sequences for each virome alongside informations about their description in viral genomes, or the fact that most of the sequences from this domain are of viral origin (identified as «viral» domains). In the second table, the 30 most retrieved GO terms are listed, with the associated number of sequences for each virome.
(DOC)

Author Contributions

Conceived and designed the experiments: TS DD. Performed the experiments: AR VR SP JC. Analyzed the data: SR FE ST. Contributed reagents/materials/analysis tools: SR FE JC. Wrote the paper: SR FE VR TS DD.

References

1. Logares R, Bråte J, Bertilsson S, Clasen JL, Shalchian-Tabrizi K, et al. (2009) Infrequent marine-freshwater transitions in the microbial world. Trends Microbiol 17: 414–422.
2. Hahn MW (2006) The microbial diversity of inland waters. Curr Opin Biotechnol 17: 256–261.
3. Lefranc M, Thénot A, Lepère C, Debroas D (2005) Genetic diversity of small eukaryotes in lakes differing by their trophic status. Appl Environ Microbiol 71: 5935–5942.
4. Short CM, Suttle CA (2005) Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. Appl Environ Microbiol 71: 480–486.
5. Chénard C, Suttle CA (2008) Phylogenetic diversity of sequences of cyanophage photosynthetic gene psba in marine and freshwaters. Appl Environ Microbiol 74: 5317–5324.
6. Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, et al. (2002) Genomic analysis of uncultured marine viral communities. Proc Natl Acad Sci U S A 99: 14250–14255.
7. Edwards RA, Rohwer F (2005) Viral metagenomics. Nat Rev Microbiol 3: 504–510.
8. Rodriguez-Brito B, Li L, Wegley L, Furlan M, Angly F, et al. (2010) Viral and microbial community dynamics in four aquatic environments. ISME J 4: 739–751.
9. López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, et al. (2009) High diversity of the viral community from an Antarctic lake. Science 326: 858–861.
10. Boucher D, Jardillier L, Debroas D (2006) Succession of bacterial community composition over two consecutive years in two aquatic systems: a natural lake and a lake-reservoir. FEMS Microbiol Ecol 55: 79–97.
11. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629–632.
12. Wommack KE, Bhavsar J, Ravel J (2008) Metagenomics: read length matters. Appl Environ Microbiol 74: 1453–1463.
13. Debroas D, Humbert J, Enault F, Bronner G, Faubladier M, et al. (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (lac du Bourget–france). Environ Microbiol 11: 2412–2424.
14. Angly F, Rodriguez-Brito B, Bangor D, McNairnie P, Breitbart M, et al. (2005) Phaccs, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. BMC Bioinformatics 6: 41.
15. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368.
16. Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, et al. (2011) Metavir: a web server dedicated to virome analysis. Bioinformatics 27: 3074–3075.
17. Carstens EB (2010) Ratification vote on taxonomic proposals to the international committee on taxonomy of viruses (2009). Arch Virol 155: 133–146.
18. Rosario K, Duffy S, Breitbart M (2009) Diverse circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol 90: 2418–2424.
19. Dorigo U, Jacquet S, Humbert J (2004) Cyanophage diversity, inferred from g20 gene analyses, in the largest natural lake in france, lake bourget. Appl Environ Microbiol 70: 1017–1022.
20. Comeau AM, Krisch HM (2008) The capsid of the t4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. Mol Biol Evol 25: 1321–1332.
21. Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64: 69–114.
22. Rodriguez-Valera F, Martin-Cuadrado A, Rodriguez-Brito B, Pasić L, Thingstad TF, et al. (2009) Explaining microbial population genomics through phage predation. Nat Rev Microbiol 7: 828–836.
23. Canchaya C, Fournous G, Brüssow H (2004) The impact of prophages on bacterial chromosomes. Mol Microbiol 53: 9–18.
24. Fuhrman JA (1999) Marine viruses and their biogeochemical and ecological effects. Nature 399: 541–548.
25. Suttle CA (2007) Marine viruses–major players in the global ecosystem. Nat Rev Microbiol 5: 801–812.
26. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, et al. (2009) The gaas metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5: e1000593.
27. Kim K, Chang H, Nam Y, Roh SW, Kim M, et al. (2008) Amplification of uncultured single-stranded DNA viruses from rice paddy soil. Appl Environ Microbiol 74: 5975–5985.
28. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4: e7370.
29. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452: 340–343.
30. Dorigo U, Fontvieille D, Humbert J (2006) Spatial variability in the abundance and composition of the free-living bacterioplankton community in the pelagic zone of lake bourget (france). FEMS Microbiol Ecol 58: 109–119.
31. Wilhelm SW, Carberry MJ, Eldridge ML, Poorvin L, Saxton MA, et al. (2006) Marine and freshwater cyanophages in a laurentian great lake: evidence from infectivity assays and molecular analyses of g20 genes. Appl Environ Microbiol 72: 4957–4963.
32. Colombet J, Robin A, Lavie L, Bettarel Y, Cauchie HM, et al. (2007) Virioplankton 'pegylation': use of peg (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. J Microbiol Methods 71: 212–219.
33. Brussaard CPD (2004) Optimization of procedures for counting viruses by flow cytometry. Appl Environ Microbiol 70: 1506–1513.
34. Personnic S, Domaizon I, Dorigo U, Berdjeb L, Jacquet S (2009) Seasonal and spatial variability of virio-, bacterio-, and picophytoplanktonic abundances in three peri-alpine lakes. Hydrobiologia 627, Number 1: 99–116.
35. Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME J 3: 1314–1317.
36. Edgar RC (2010) Search and clustering orders of magnitude faster than blast. Bioinformatics 26: 2460–2461.
37. Reyes A, Haynes M, Hanson N, Angly FE, Heath AC, et al. (2010) Viruses in the faecal microbiota of monozygotic twins and their mothers. Nature 466: 334–338.
38. Martín-Cuadrado A, López-García P, Alba J, Moreira D, Monticelli L, et al. (2007) Metagenomics of the deep mediterranean, a warm bathypelagic habitat. PLoS One 2: e914.
39. Suzuki R, Shimodaira H (2006) Pvclust: an r package for assessing the uncertainty in hierarchical clustering. Bioinformatics 22: 1540–1542.
40. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403–410.
41. Cole JR, Wang Q, Cardenas E, Fish J, Chai B, et al. (2009) The ribosomal database project: improved alignments and new tools for rrna analysis. Nucleic Acids Res 37: D141–5.
42. Li L, Kapoor A, Slikas B, Bamidele OS, Wang C, et al. (2010) Multiple diverse circoviruses infect farm animals and are commonly found in human and chimpanzee feces. J Virol 84: 1674–1682.
43. Sullivan MB, Krastins B, Hughes JL, Kelly L, Chase M, et al. (2009) The genome and structural proteome of an ocean siphovirus: a new window into the cyanobacterial 'mobilome'. Environ Microbiol 11: 2935–2951.
44. Filée J, Comeau AM, Suttle CA, Krisch HM (2006) T4-type bacteriophages: ubiquitous components of the dark matter of the biosphere]. Med Sci (Paris) 22: 111–112.
45. Huang X, Madan A (1999) Cap3: a DNA sequence assembly program. Genome Res 9: 868–877.
46. Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56: 564–577.
47. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
48. Letunic I, Bork P (2007) Interactive tree of life (itol): an online tool for phylogenetic tree display and annotation. Bioinformatics 23: 127–128.

La spécificité des communautés virales lacustres, reflet de l'importance de la salinité

Ces premiers résultats obtenus sur des échantillons de lacs d'eau douce en climat tempéré associés aux études précédemment publiées sur les communautés virales des milieux aquatiques semblent témoigner d'une forte similarité des viromes issus de ce type d'écosystème. Cette tendance est d'autant plus marquante que les milieux d'eau douce étudiés par métagénomique virale jusqu'ici sont de nature très différente, constitués de bassins piscicoles en Amérique du Nord, d'un lac dont la surface est gelé neuf mois sur douze (le Lac Limnopolar), et des lacs Pavin et Bourget qui, s'ils sont tous deux européens, diffèrent par leur taille, leur volume et leur statut trophique. A l'inverse, plusieurs viromes de milieux proches géographiquement de ces environnements d'eau douce sont inclus dans l'analyse, et semblent génétiquement plus distants. Ainsi, cette similarité entre viromes d'eau douce reflète vraisemblablement une forte structuration des communautés virales par la salinité, pouvant se faire de manière directe *via* les contraintes imposées par ces concentrations en sel sur les capsides virales, ou plus vraisemblablement de manière indirecte de par les différences entre les communautés d'hôtes potentiels. Ces résultats sont en accord avec les hypothèses généralement admises quant aux possibilités de dispersion des virus et l'existence de biogéographie : il n'existe visiblement pas de limite de dispersion pour la majeure partie des capsides virales.

Au-delà des similarités en terme de contenu génétique, la richesse en gènes des communautés virales pourrait aussi différer entre les environnements, avec à nouveau une structuration par la salinité. Les viromes issus d'environnements marins comportent notamment un nombre de séquences différentes plus importante que les autres milieux aquatiques échantillonnés, notamment les milieux d'eau douce et les milieux hypersalins.

La distinction entre les communautés virales d'eau douce et les autres milieux aquatiques semble se déterminer principalement au niveau de l'adaptation d'un nombre restreint de groupes, ce qui impliquerait comme postulé par Logares et collaborateurs que les transitions entre les différents écosystèmes existent mais sont limitées à des événements ponctuels (Logares *et al.*, 2009). Ce faible taux de transition correspond à l'observation de clades réunissant un ensemble de virus lacustres distincts, mais dont l'origine est commune. Il est à noter que les principaux virus retrouvés lors de cet étude sont différents des virus décrits et observés dans ces mêmes lacs lors d'études antérieures (Colombet, 2008). En effet, les petits virus à ADN simple brin comme les *Microviridae* et les *Circoviridae*, de par la faible taille de leur capside et de leur génome, n'avaient ainsi jamais été véritablement pris en compte lors des études en microscopie électronique ou PFGE. Toutefois, beaucoup de

viromes issus de différents types d'échantillons semblent dominés par ce type de virus (Angly *et al.*, 2009; López-Bueno *et al.*, 2009; Rosario *et al.*, 2009a), ce qui laisse craindre l'existence d'un biais systématique introduit lors de la préparation des viromes. Il est en effet possible que l'étape d'amplification aléatoire de génome complet, indispensable pour disposer d'une quantité suffisante de matériel génétique, amplifie préférentiellement les petites séquences d'ADN circulaires. Un travail complémentaire reste ainsi à effectuer afin de déterminer la part exacte prise par ce type de virus dans la communauté virale aquatique.

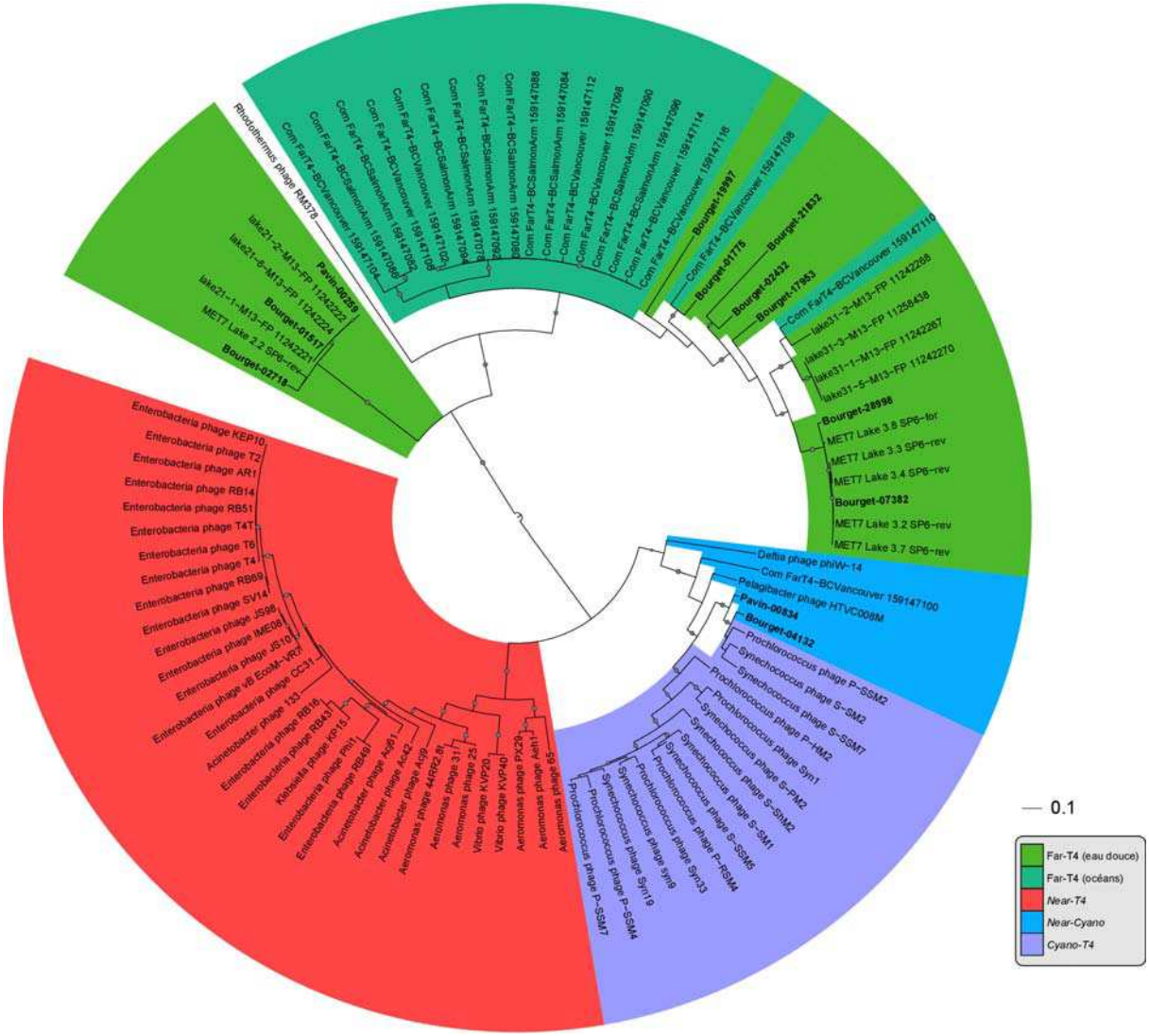


Figure III.1 : Arbre basé sur la protéine majeure de capside comprenant les trois grands groupes associés aux phages de type T4 : Near-T4, Cyano-T4 et Far-T4. *Les séquences issues de l'assemblage des viromes sont surlignées en gras, et les nœuds dont le support de bootstraps est supérieur à 50 sont notés par un cercle. Le groupe Near-T4 comprend le phage T4 en lui même et les phages les plus proches infectant différents types de gamma-proteobactéries. Les Cyano-T4 comprennent majoritairement des cyanophages de type T4, certaines séquences métagénomiques associées, ainsi que des phages infectant des alpha et beta-proteobacteria, identifiés comme "Near-Cyano". Enfin, le groupe des Far-T4 comprend différentes séquences amplifiées ou issues de viromes, ainsi qu'un phage de Bacteroidetes.*

Le clade Far-T4, détecté dans les deux viromes lacustres, était quant à lui considéré comme uniquement marin sur la base de résultats de clonage-séquençage. Un travail complémentaire sur ces virus, basé sur un assemblage des séquences, a par la suite révélé que les amorces PCR basées sur les séquences marines ne pouvaient pas amplifier les exemplaires lacustres de ces gènes. Une série d'amorces a ainsi été générée, et utilisée pour caractériser plus précisément la diversité de ces virus au sein de différents lacs (Figure III.1). Ces premiers résultats associés aux descriptions récentes de phages de type T4 infectant certaines protéobactéries parmi les plus abondantes en milieu marin (Zhao *et al.*, 2013) montrent qu'une partie de la diversité au sein de cette famille de bactériophages reste à caractériser. Le spectre d'hôte associé s'étend ainsi des différents types de *Proteobacteria* aux *Cyanobacteria* et *Bacteroidetes*, soit les clades bactériens les plus fréquemment retrouvés en milieux aquatiques. Il est donc fort probable que les phages de type T4 constituent des acteurs essentiels des milieux aquatiques en général, et des écosystèmes lacustres en particulier.

Évolution des communautés virales le long d'un gradient de salinité

Si, comme semblent l'indiquer les résultats des études précédentes de métagénomique virale, la salinité est effectivement l'élément structurant principal des communautés virales, alors des écosystèmes même proches géographiquement devraient présenter des communautés virales distinctes si leur concentration en sel est différente. Afin de vérifier cette hypothèse, une série de viromes a été réalisée le long d'un gradient de salinité, dans le cadre du programme Archevir (programme FRB, responsable : Télesphore Sime-Ngando), qui visait à caractériser les relations entre les virus d'archées et leurs hôtes dans les milieux hypersalins. Plusieurs stratégies ont ainsi été mises en place, dont des approches de culture et d'isolement, une caractérisation des communautés microbiennes procaryotes et eucaryotes par l'analyse d'amplicons d'ADNr (16S et 18S), et la recherche de gènes fonctionnels spécifiques des archées. Il sera ici question uniquement de la partie cherchant à décrire par approche métagénomique les communautés virales, leur composition, leur potentiel fonctionnel, et leur relations avec les autres communautés virales hypersalines décrites jusqu'alors.

Le lieu principal d'échantillonnage, Ngallou, est un estuaire inverse situé au sud de Dakar au Senegal, ce qui provoque dans un espace restreint la formation de bassins avec une salinité croissante (du niveau de la mer à des concentrations proches de la saturation). Un échantillon supplémentaire a été prélevé au niveau du lac Rose, lac hypersalin, pour lequel une analyse de la communauté virale avait déjà été effectuée au sein du laboratoire (Sime-Ngando *et al.*, 2010). Ce projet précédent avait révélé l'existence d'une diversité de structures

de capsides exceptionnelle, avec la description de plusieurs nouvelles formes, comme des capsides en épi de maïs ou en crochets (Figure In.2).

Les viromes du projet Archevir ont de plus bénéficié de l'utilisation du séquençage HiSeq 2000 d'Illumina, qui a permis de générer plusieurs dizaines de millions de séquences pour chaque échantillon. Si ce nombre de séquence est beaucoup trop important pour une analyse directe, il procure cependant un taux de couverture des génomes suffisant pour effectuer un assemblage, et ainsi analyser de larges fragments génomiques, voire des génomes complets.

Enfin, l'étude de ces viromes s'est appuyé sur plusieurs résultats de métagénomique virale appliquée aux milieux hypersalins. En 2010, Rodriguez-Brito et collaborateurs ont ainsi analysé plusieurs bassins hypersalins californiens, et montré une séparation des communautés en fonction de la salinité, ainsi qu'une relative stabilité de ces communautés dans le temps (Rodriguez-Brito *et al.*, 2010). En 2012, deux autres études de viromes hypersalins ont été publiées. L'une concernait le suivi sur trois ans des communautés virales du Lac Tyrrell, lac hypersalin situé en Australie, où à l'inverse des résultats précédents, les auteurs ont cette fois mis en avant des modifications importantes entre les différents points d'échantillonnage au cours du temps, et une absence de similarité entre les séquences issues du lac Tyrrell et celles échantillonnées en Amérique du Nord (Emerson *et al.*, 2012). En parallèle, une étude basée sur des fosmides (fragments d'environ 40 kb) issus d'un échantillon d'une saline espagnole a permis de décrire de nouveaux fragments génomiques de *Caudovirales* adaptés à ces milieux hypersalins, infectant majoritairement des archées (Garcia-Heredia *et al.*, 2012).

Dans ce contexte, les viromes du programme Archevir devaient apporter des réponses quant à la répartition des virus au niveau mondial, et en particulier au sujet des similarités entre les différentes communautés virales hypersalines autour du globe. De plus, l'étude de fragments génomiques devrait permettre de déterminer le(s) niveau(x) et le(s) moyen(s) de différenciation génétique et génomique de ces virus hypersalins par rapport aux autres milieux aquatiques.

Article V

Meta-analysis of metagenomic data shows that halophilic viral pan-genome is consistent across time and space

Roux Simon[1,2], Enault François[1,2], Ravet Viviane[1,2], Vellet Agnès[1,2], Bettarel Yvan[3,] Auguet Jean-Christophe[4], Bouvier Thierry[3], Lucas-Staat Soizick[5], Forterre Patrick[5,6], Prangishvili David[5], Debroas Didier[*1,2], Sime-Ngando Télésphore[1,2]

[1] Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", Clermont-Ferrand , France [2] CNRS UMR 6023, LMGE, Aubière, France [3] UMR 5119 ECOSYM, CNRS, IRD, Ifremer, Université Montpellier 1 [4] IPREM-EEM UMR5254, University of Pau [5] Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, Paris, France [6] Laboratoire de Biologie Moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Orsay, France

Manuscrit en préparation

Matériel supplémentaire : Annexe A.6

Meta-analysis of metagenomic data shows that halophilic viral pan-genome is consistent across time and space

Roux Simon[1,2], Enault François[1,2], Ravet Viviane[1,2], Vellet Agnès[1,2], Bettarel Yvan[3,] Auguet Jean-Christophe[4], Bouvier Thierry[3], Lucas-Staat Soizick[5], Forterre Patrick[5,6], Prangishvili David[5], Debroas Didier[*1,2], Sime-Ngando Télésphore[1,2]

[1] Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", Clermont-Ferrand , France [2] CNRS UMR 6023, LMGE, Aubière, France [3] UMR 5119 ECOSYM, CNRS, IRD, Ifremer, Université Montpellier 1 [4] IPREM-EEM UMR5254, University of Pau [5] Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, Paris, France [6] Laboratoire de Biologie Moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Orsay, France

Keywords : Hyperhalophile, viruses, metagenomics

Abstract

Microbial communities from hypersaline ponds represent unique assemblies among the aquatic environments' micro-organisms, most of them being dominated by halophilic members of the *Archaea* domain ('*Haloarchaea*'). These organisms adapted to life in hypersaline ponds are considered to be specific of such extreme conditions, and the associated viral communities have accordingly been shown to display remarkable features, such as highly unusual morphologies or genome organization. Here, we studied the diversity of viral communities from a salinity gradient (8 to 36 % of salinity) through a high-depth sequencing metagenomics approach. The different viral communities could thus be characterized on different scales, from the community to the genotype level. Overall, hyperhalophilic viral communities appear to be highly specific and consistent among the different hypersaline environments sampled on Earth, both at the community and the clade scale. More specifically, a limited number of clades within the main retrieved group of Caudovirales seem to be adapted to these hypersaline environments, and were retrieved in all hyperalophilic samples from three different continents. More than genetic exchange, analysis of genomic fragments tends to indicate that these similarities are indeed due to related lineages with a world-wide distribution. Conversely, other recently described hyperhalophilic viruses (Salterprovirus) display a far greater rate of gene transfer and recombination within these hypersaline ponds, both between different kind of viruses and between viruses and mobile genetic elements. Thus, hypersaline viral communities around the world appear to form a genetically consistent community, within which genetic exchange can occur at multiple scales and with a wide range of genetic templates. Considering the amount of uncharacterized sequences in these different metagenomes, such environments are very likely to contain some new and interesting way of dealing with extreme salt concentration, which associated genes would be transferred and disperesed through halophilic viral genomes.

Introduction

Hypersaline environments are defined as environments with a salinity above that of seawater up to salt saturation. Such environments are found all around the world, especially in hot and dry regions either in natural salt lakes or in crystallizer ponds from solar salterns. These ecosystems harbor a low species richness and high density of prokaryotic cells, often exceeding 10^7 per milliliter (Benlloch et al., 2002; Bettarel et al., 2011; Casamayor et al., 2002). Even if members of all three domains of life have been identified in such ecosystems, halophilic *Archaea* are the most abundant organisms, their dominance usually increasing with salt concentration (Benlloch et al., 2002; Casamayor et al., 2002; Sime-Ngando et al., 2010). Among these halophilic *Archaea*, most representatives belong to the family *Halobacteriaceae.* The most famous member of this family of *Archaea* adapted to the halophilic conditions is *Haloquatrum Walsbyi* (Bolhuis, Poele, & Rodriguez-Valera, 2004) for its square shaped cells, and because it dominates the microbial populations of salt lakes (Oh, Porter, Russ, Burns, & Dyall-Smith, 2010).

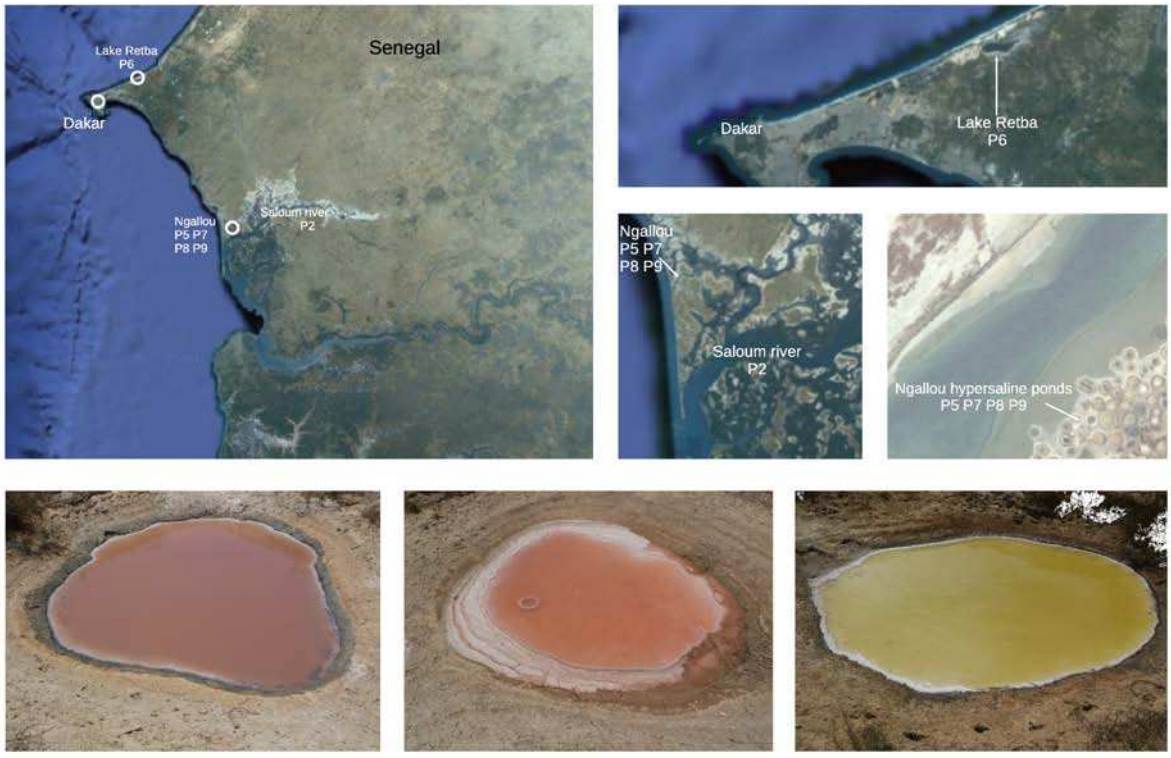
Beside these cellular organisms, hypersaline systems harbor the highest density of virus-like particles reported for aquatic systems (up to 10^9 virus-like particles per ml, (Baxter et al., 2011) and virus-to-cell ratios considerably greater than in most environments (Laybourn-Parry, Hofer, & Sommaruga, 2001). Beyond their impact on the microbial genome evolution through horizontal gene transfer (Rohwer, Prangishvili, & Lindell, 2009), their role in the regulation of microbial population is of special importance in these hypersaline environments where bacteriovory is scarcely detected at high salt concentration (Pedrós-Alió et al., 2000).

Observations with electronic microscopy and flow cytometry on hypersaline samples demonstrated a great abundance and morphologic diversity of viruses in such ecosystems (Bettarel et al., 2011; Schapira et al., 2008; Sime-Ngando et al., 2010). Isolation and cultivation of halophilic microorganisms led to the description of a handful of their associated viruses. To date, 13 haloviruses infecting *Archaea* from the *Halobacteria* class are completely sequenced. Described morphotypes of these viruses range from tailed and tailless icosahedral viruses to pleomorphic viruses lacking well-defined shapes (Pina, Bize, Forterre, & Prangishvili, 2011; Roine & Oksanen, 2011). Though the morphotype diversity of these viruses do not reach the one of crenarcheal viruses, they display some great variability, especially in terms of genomic composition, and even between related viruses such as the

two members of Salterprovirus group : His1 and His2 (Roine et al., 2010). Additionally, only a few haloviruses infecting bacteria such as *Salisaeta* icosahedral phage 1 (SSIP-1 ; (Aalto et al., 2012)) are described, since bacteria from hypersaline environments remain especially difficult to cultivate in the lab. Overall, only 400 of the 4334 proteins available in the RefseqVirus database originate from haloviruses.

Metagenomics approaches are able to provide further insights on the viral communities genetic pool, and were consequently used to characterize different hypersaline viral communities (Boujelben et al., 2012; Emerson et al., 2012; Garcia-Heredia et al., 2012; Rodriguez-Brito et al., 2010; Santos, Yarza, Parro, Briones, & Antón, 2010; Sime-Ngando et al., 2010). These analyses, based on limited number of fosmids or on short reads, all came to the conclusion that hypersaline viral communities were both mostly uncharacterized and specific to these ecosystems. Indeed, the vast majority of the metagenomic sequences did not display any similarity with known viral genomes, the only sequences with a recognizable homologue being similar to the few viruses isolated from hypersaline samples. Yet, unknown sequences seemed to present some significant similarities throughout the different hypersaline samples (Boujelben et al., 2012; Sime-Ngando et al., 2010). More recently, sets of large genomic fragments were generated from the viral fraction of a salt lake (Lake Tyrrell in Australia; (Emerson et al., 2012)) and from solar salterns ponds (Santa Pola pond in Spain; (Garcia-Heredia et al., 2012)). These genomic fragments, generated using respectively high-depth sequencing (Illumina Hi-Seq2000) and complete sequencing of fosmids, are especially interesting as complete protein sequences can be predicted and genomic architecture can be analyzed (*i.e.* gene content and gene order). These two studies came to contradictory conclusions regarding the geographical and temporal variability of the viral communities. Most of the viruses assembled from the Lake Tyrrell samples appeared to have a very dynamic pattern of presence-absence throughout the time series collected, and did not presented any similarities with previous hypersaline viromes. On the contrary, Santa Pola fosmids were described as very similar to previous datasets sampled both from the same location and from different hypersaline ponds (especially to the San Diego viromes (Rodriguez-Brito et al., 2010)). Eventually, the actual diversity of hypersaline viral communities, their distribution around the globe and their links to lower salinity mediums are still to be assessed.

Here, we present the analysis of 6 viromes generated from a salinity gradient (8 to 36 % salinity). Using high-depth sequencing, we were able to assemble large genomic fragments, and putative complete genomes of various sizes from all samples along the salinity



Samples chemical properties					Initial assembly					Filtered contigs			
Sample	Salinity	Cl^- (mg/L)	Na^+ (mg/L)	Mg^{++} (mg/L)	Number of Reads	Number of contigs	Contigs average size (nt)	Contigs maximum size (nt)	Number of proteins	Number of contigs	Contigs average size (nt)	Contigs maximum size (nt)	Number of proteins
P2	8.0%	1410	856	145	44,532,000	89,754	614.19	60,956	133,290	7,302	2125.52	39,349	33,025
P5	26.0%	4680	3637	405	46,905,300	44,320	605.66	44,814	63,725	3,018	2405.89	44,814	14,476
P6	29.0%	3840	2063	1058	53,825,100	26,592	335.88	38,687	31,234	884	2433.75	38,687	4,567
P7	29.8%	5240	3507	952	42,769,700	27,774	468.80	48,623	36,093	1,312	2494.48	33,039	6,559
P8	35.0%	5400	3367	1025	51,795,500	41,402	605.41	57,903	62,332	3,212	2408.65	57,903	16,277
P9	36.0%	5600	1735	2195	62,616,100	17,173	757.81	45,851	24,414	529	3577.20	39,266	2,979

Figure 1: Location and description of the different samples along the salinity gradient. For each sampling point, physico-chemical, sequencing and assembly parameters are indicated.

gradient. These different communities could thus be characterized from a community scale to a viral "strain" level, and compared to previously reported datasets through a meta-analysis.

Results

Six metagenomes are analyzed in this study, sampled from six locations in Senegal at the end of the dry season (May 2011), with salinities ranging from 8 to 36 % (Fig 1). The P2 sample was collected in the inverse estuary of Saloum, with a measured salinity of 8% *i.e.* about twice the seawater salinity. P5, P7, P8 and P9 were collected in individual small marine solar salterns that consist of shallow ponds where seawater is concentrated until sodium chloride is precipitated. P6 was sampled in Lake Retba, a natural salt lake. The physico-chemical properties of the different samples were rather uniform, with the main salt components being Na^+ and Cl^- for all samples except for P9, which main cation is Mg^{++}. Viral capsids were precipitated for each sample, and the associated DNA pool was randomly amplified and sequenced with Hi-Seq2000 (Illumina), leading to 42 to 60 millions paired-end 100bp sequences available for each sample, subsequently assembled in several thousands contigs for each point.

Changes within the viral community along the salinity gradient

A gene prediction was applied to all our contigs, in order to get a set of predicted genes for each sample. Affiliation results of the six datasets are consistent with viral metagenomes analysis published so far, as more than half of the predicted proteins (62 to 82 %) is not similar to any sequence of the NR database, and identified sequences are roughly equally divided between viruses on one side and *Bacteria* or *Archaea* on the other (Fig 2A).

In order to dispose of more accurate affiliation of the viral community, virome contigs were compared to complete viral genomes from RefseqVirus database (Fig. 2B). On overall, members of the Caudovirales families (*Myo*, *Sipho* and *Podoviridae*) are the most retrieved viruses regardless the salinity level. As could be expected, the relative number of viruses known to infect *Archaea* increases with the salinity level from P2 to P8, due to the detection of several viruses specific of hypersaline environments like members of the Salterprovirus genus (Bath, Cukalac, Porter, & Dyall-Smith, 2006), Halovirus HF1 (Tang, Nuttall, & Dyall-smith, 2004) or Halovirus HVTV1 (Pietilä et al., 2013). Yet the ratio of these hyperhalophilic viruses is lower in the P9 sample, which could indicate that the viral community of this very peculiar sample (salinity close to the saturation level and dominated by Mg^{2+} rather than Na^+ cations, Fig 1) could be specific.

Beta diversity of viruses driven by salinity

In order to compare viromes both within this study and with previously published datasets (Table S1), we selected 42 aquatic viromes of different salinity : freshwater, seawater (4 %), low and medium hypersaline (6-14 %) and high hypersaline samples (25-36 %).

Overall, viromes appear to be gathered according to the samples salinity (Fig. 3). Namely, the P2 sample, which salinity was measured at 8 %, is gathered with medium hypersaline viromes from San Diego solar saltern (12 – 14 % salinity; CA, USA; Rodriguez-Brito et al., 2009). Consistently, all high hypersaline viromes (>25%) are gathered : P5-6-7-8 and 9 from Senegal, 7 viromes from Lake Tyrrell samples (Australia; (Emerson et al., 2012)) and the high hypersaline samples from San Diego solar saltern (CA, USA; (Rodriguez-Brito et al., 2010)). Positive correlation found between MDS coordinates (especially axis 1) and salinity values for each sample confirmed that salinity was a strong discriminating factor (p-value=0.001).

Additionally, genetic distances seems higher within high salinity viromes than between seawater or low-medium salinity viromes. Samples P6 and P9 are clearly separated from other hypersaline ponds, even when compared to samples from the same geographical location (P5-7-8). Interestingly, these two viromes (P6 and P9) originate from the two samples with the highest Mg^{2+} to Na^+ ratio. The Lake Tyrrell samples are on the other hand closely tight together despite spanning more than 3 years of sampling, two different sampling points within the lake, and salinity level ranging from 24 to 36 (Emerson et al., 2012). For all these samples, Na^+ was the most retrieved cation. These different observations tends to indicate that geochemical properties could drive the distribution of viral communities more strongly than geographical or temporal distance, and that within hypersaline viral comunities, relative abundance of the different chemical compounds could be more important than small variations in the overall quantity of dissolved salt.

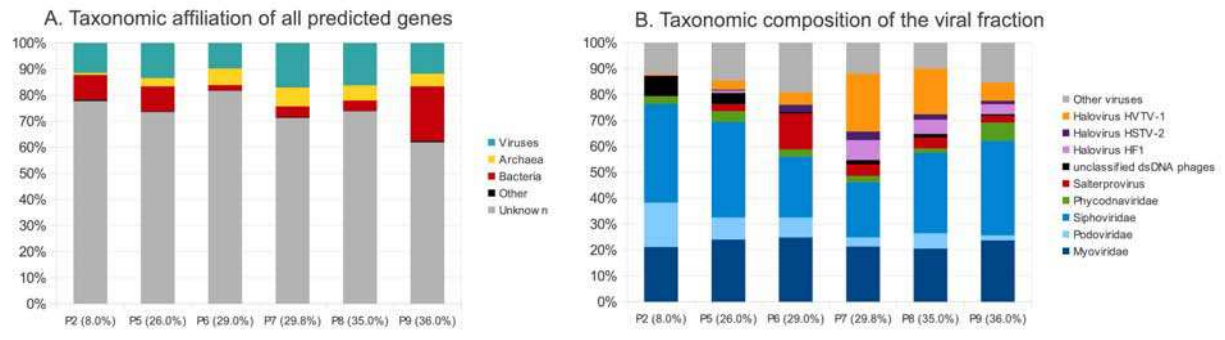


Figure 2 : Taxonomic composition of Senegal viromes. (A) Global taxonomic composition, based on the affiliation of predicted ORFs to the kingdom level from a best BLAST hit against NR. (B) Viral taxonomic composition at the viral family level assessed for all virus-associated contigs.

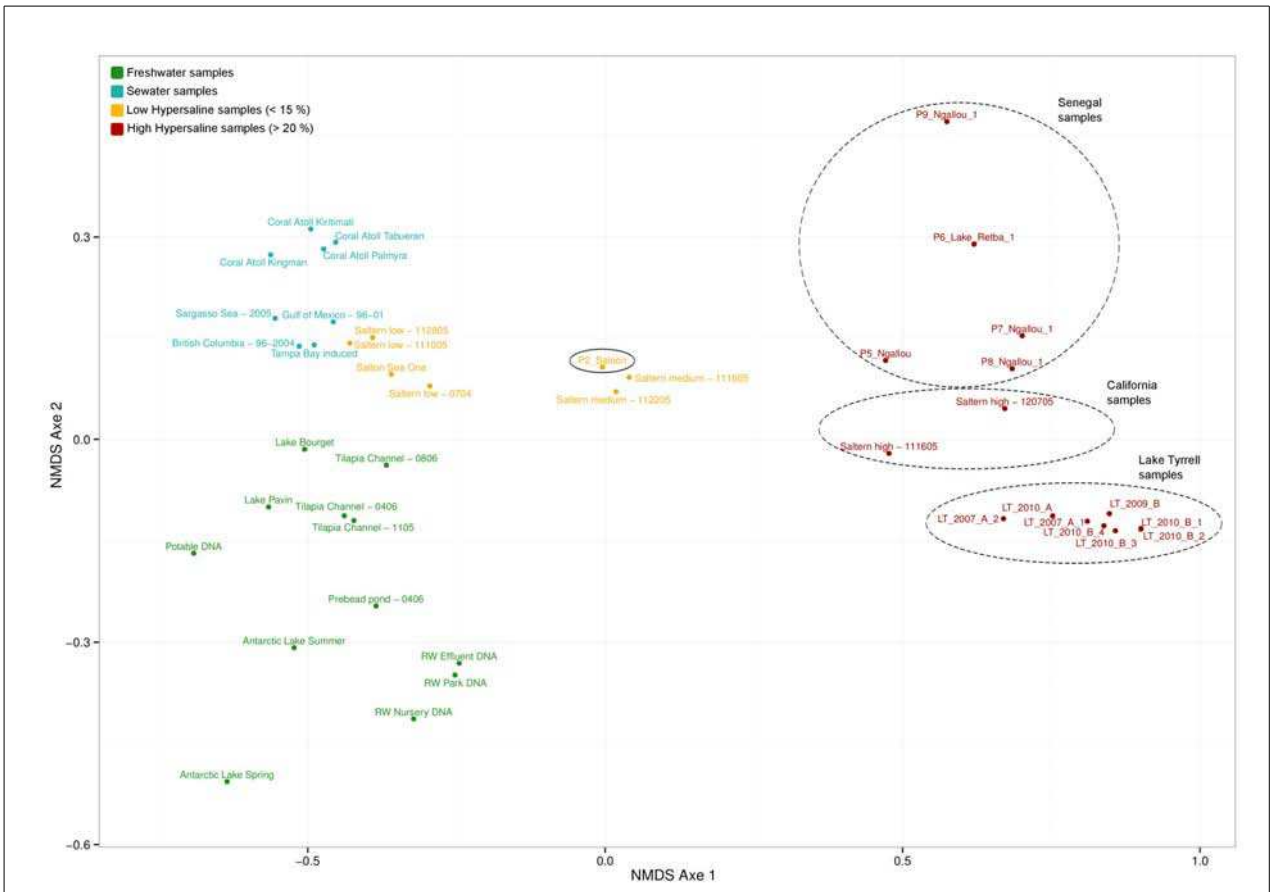


Figure 3 : Global comparison of aquatic viromes. Reciprocal BLAST were computed for viromes sub-samples (50,000 sequences of 100bp), and results from these BLAST comparison were used in a NMDS to display viromes according to their average genetic similarity. All sub-samples were taken from raw reads. Viromes are colored according to the salinity level of the sample, and the three groups of high hypersaline samples are framed (dash strokes). The sample of intermediate salinity from Ngallou (P2) is circled.

Halophilic Caudovirales Group are widespread around the globe

Caudovirales are usually the most retrieved DNA viruses in the environment, and hypersaline ponds are no exception (Fig 2B). Within contigs affiliated to the Caudovirales order, half of them are affiliated to the *Siphoviridae* family, about a third to the *Myoviridae* family, the rest being associated with *Podoviridae*, less frequently retrieved in all samples (Table S2). Within Caudovirales, *Natrialba* phage PhiCH1 and Archaeal BJ1 virus stand out as consistently retrieved. Halovirus HF1, an archaeal virus considered as "unclassified dsDNA virus" but which displays a head-tail morphology (Tang et al., 2004) and which genome contains several marker genes of the Caudovirales order (for instance the large terminase subunit), is also frequently retrieved.

In order to further examine the diversity of hyperhalophilic *Caudovirales,* a phylogenetic tree was built from a *Caudovirales* specific marker gene (TerL, coding for the large subunit of the terminase), including sequences from both complete viruses and metagenomic datasets (Fig 4). This tree makes it possible to distinguish between low hypersaline Senegal sample (P2) and all the other high hypersaline samples. Indeed, sequences from the low hypersaline sample P2 are widely distributed throughout the tree, whereas 74% of the 175 environmental sequences from the high hypersaline samples are gathered in a limited number of groups. We thus selected all monophyletic groups with a bootstrap support greater than 80% and containing more than five hypersaline metagenomics sequences, and identified them as Hypersaline Caudovirales Group (HCG) 1 to 6 (highlighted in orange on Fig 4).

These six groups gather sequences from at least two locations, and four groups (HCG1, 2, 4 and 6) are actually formed by sequences from hypersaline samples of the three continents studied here (Santa Pola in Europe, Lake Tyrrell in Australia, and Ngallou in Africa). Only one viral reference sequence is contained in one of these hypersaline group (HCG 1) : Archaeal virus BJ1, a virus isolated in a hypersaline lake in Inner Mongolia with a Haloarchaeal host (Pagaling et al., 2007). Several TerL sequences from cellular genomes are also retrieved in these HCG, including four sequences from *Haloarchaea*, namely *Haloterrigena turkmenica* and *Natronobacterium gregoryi* for HCG 1 and *Halorubrum lacusprofundi* and *Halalkalicoccus jeotgali* for HCG 2. Yet, no other gene of viral origin could be detected in the neighborhood of the TerL gene in these cellular genomes, preventing the detection of any integrated prophage. Still, based on the recovery of such related TerL homologs, it is tempting to speculate that members of HCG1 and HCG2 infect *Haloaarchaea*. Two bacterial sequences are also retrieved within HCG3, originating from genomic sequences of

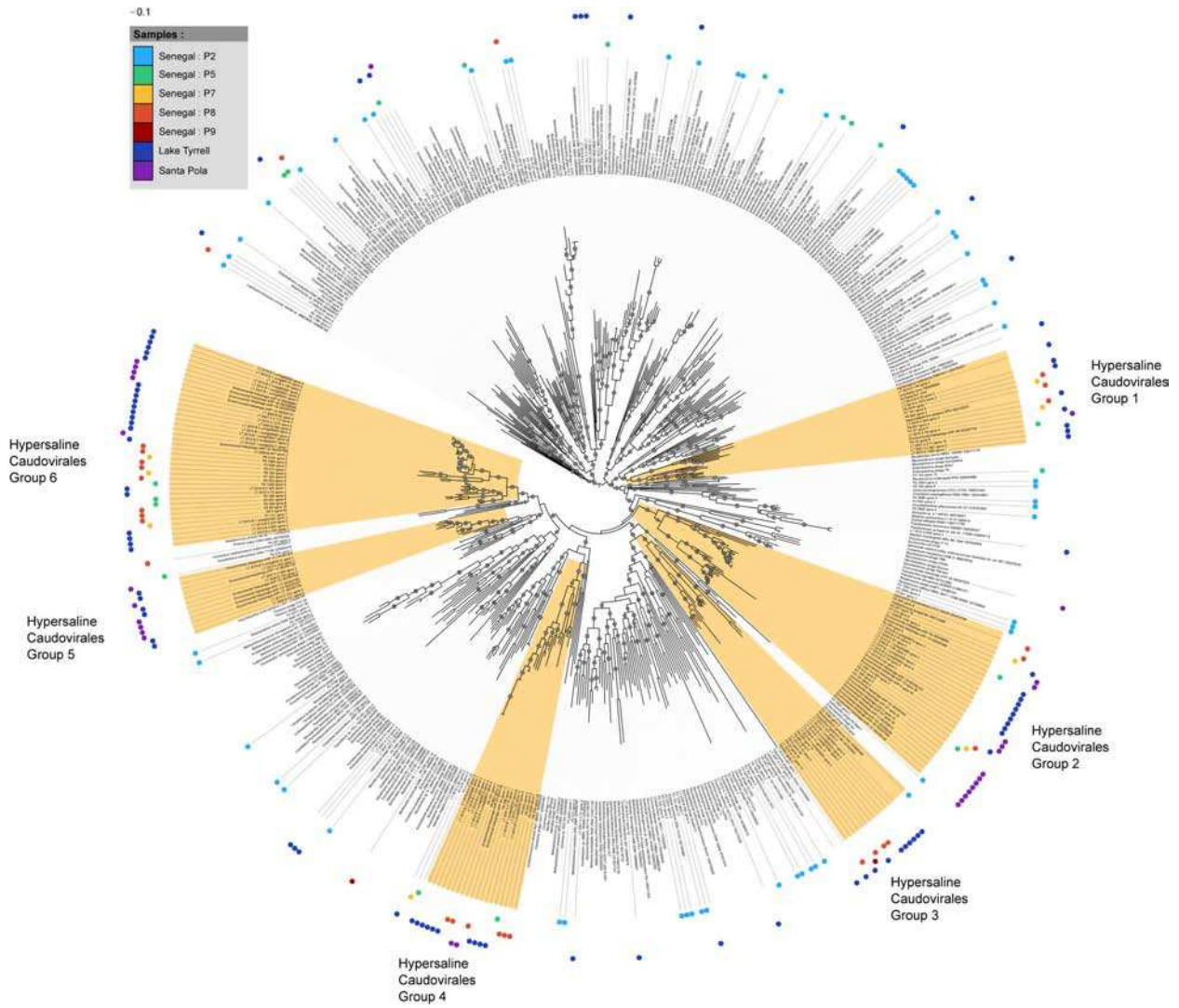


Figure 4 : Phylogenetic tree of Caudovirales computed from the large subunit terminase. Hypersaline Caudovirales groups are highlighted in orange. The origin of each virome sequence is indicated with a colored dot in front of the contig name. Boostraps scores greater than 0.80 are indicated with a gray dot.

Alpha-proteobacteria (*Rhodospirillum centenum* and *Azospirillum lipoferum*), pointing toward a putative bacterial host for the members of this group, though no prophages could be detected in that case either.

In addition to this observation of genetic diversity, several long contigs were available for *Caudovirales,* especially interesting to study the gene composition, order and conservation within this family. We thus selected all contigs of more than 10 Kb in the defined Hyperhalophilic Caudovirales Groups for genomic analysis (Table S3). First, these 41 sequences were compared through a reciprocal BLASTp of all predicted proteins. The clustering computed from this whole contig comparison is consistent with the phylogenetic tree (Fig S1), indicating that genomic fragments from the same group on the TerL tree are indeed more similar throughout their whole sequence to each other than to other groups. A conservation of the gene order between contigs could be observed within each group, even for contigs sampled from distant location (Fig 5). The large subunit terminase gene (TerL) is often associated with conserved structure gene (*e.g.* portal or capsid protein ; all groups except HCG5), and more scarcely with replication-associated genes (only for groups HCG3 and HCG6, Fig 5). Interestingly, several uncharacterized genes in these modules are conserved throughout the different locations, therefore likely to code for important functions. These genes could thus be candidates of choice for subsequent targeted functional analysis.

Finally, we took advantage of the high depth of Hi-Seq Illumina sequencing to assess the average variability represented by each contig (Table S4). The ratio of mismatch found when mapping reads against these contigs was similar over all contigs and quite low (0.25 – 0.53 %), which indicates that the different contigs do not stand for a very large and diverse population, but rather for a single genotype.

Salterprovirus-like genes are retrieved in a large range of genomes

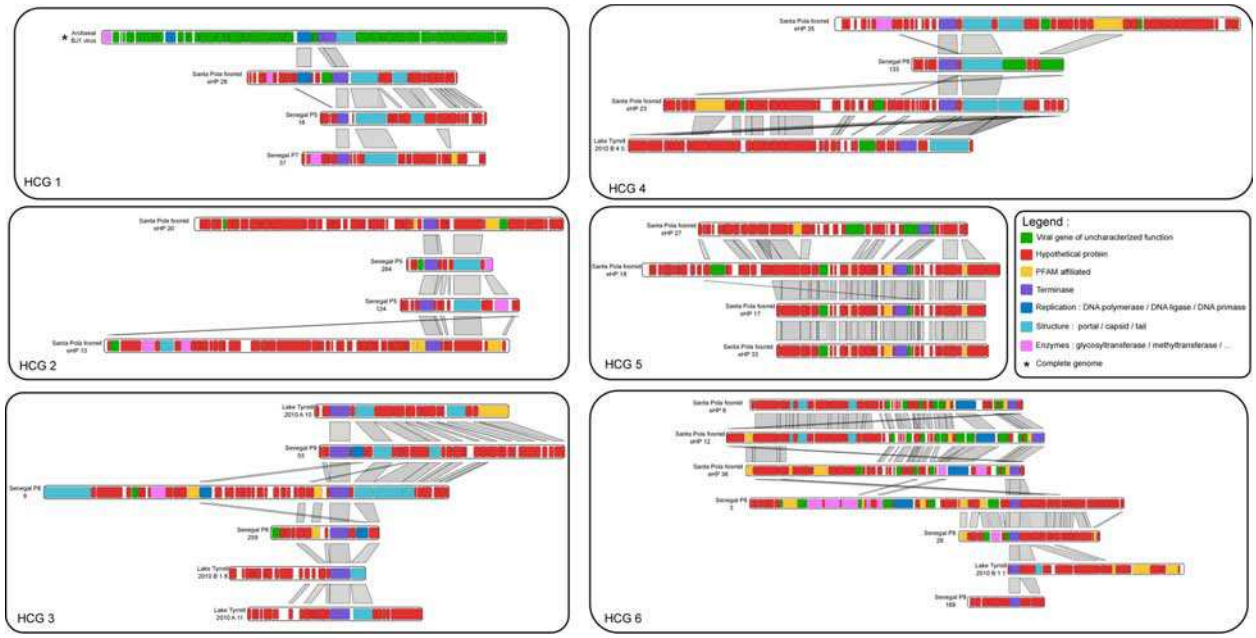


Figure 5 : Gene order conservation throughout the different Hypersaline Caudovirales Groups. Genes are colored according to their functional affiliation.

Alongside Caudovirales, several viral families specifically linked to hypersaline ponds were retrieved in the high hypersaline viromes. Notably, several large contigs were found to be similar to Salterproviruses, small viruses with lemon-shaped, enveloped virions, isolated from hypersaline environments and infecting hyperhalophilic *Archaea (Bath et al., 2006)*. Accordingly, virions displaying the same specific lemon shape were observed in Senegal hypersaline samples (Fig S2). The two Salterprovirus genomes currently available (named His1 and His2), are of similar size and nature (linear, around 15kb, dsDNA), but display only one common gene (thought to be involved in genome replication). Contigs similar to these viruses were thus further investigated in order to assess the distribution of Salterprovirus genes in different hypersaline ponds, as well as the genomic context in which such genes are retrieved.

Even if several contigs are affiliated to Salterprovirus His2 for most of the assembled hypersaline viromes (Table S2), one single gene (His2_gp07, a putative argynil tRNA synthetase) is actually retrieved in large (>5kb) or circular contigs (Table S5). The existence of viral copies of aminoacyl tRNA synthetases (aaRS) genes was first attested in Mimivirus genome (Raoult et al., 2004), and subsequently confirmed by additional genome sequences of large eukaryotic (Arslan, Legendre, Seltzer, Abergel, & Claverie, 2011; Fischer, Allen, Wilson, & Suttle, 2010) and bacterial viruses (Hendrix, 2009). In

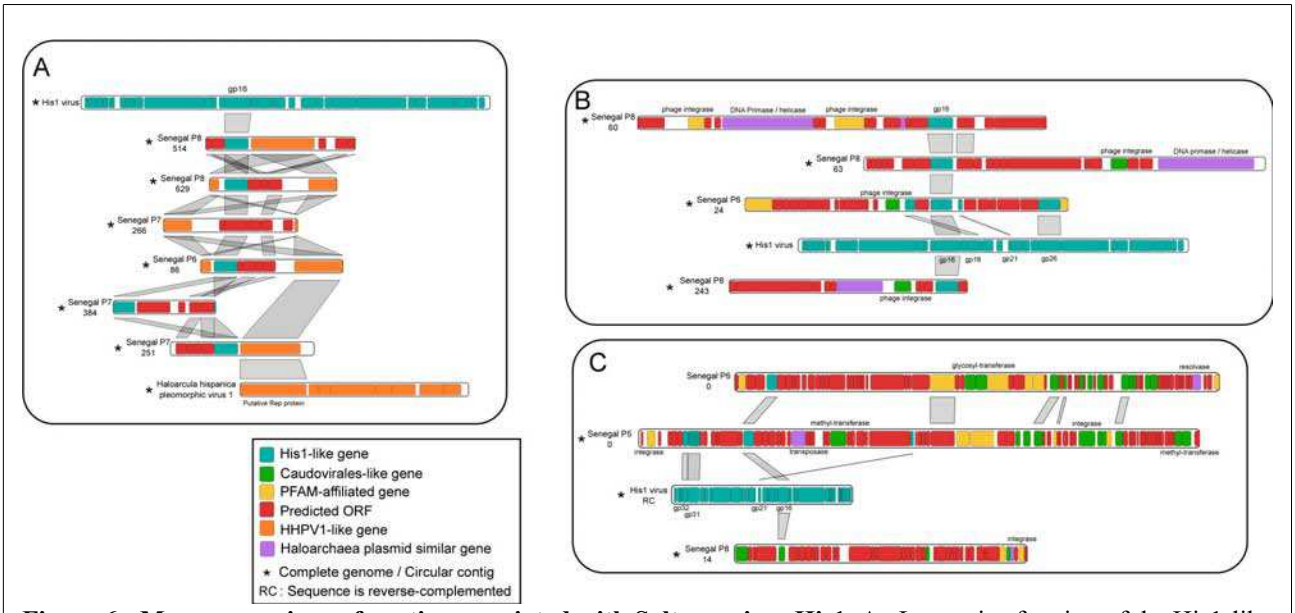


Figure 6 : Map comparison of contigs associated with Salteprovirus His1. A : Large size fraction of the His1-like small circular contigs. B : Putative complete His1-like genomes. C : Genomic fragments including both His1-like genes and large dsDNA viruses associated sequences.

hypersaline contigs, this His2-like aaRS gene is retrieved in contigs associated to Archaeal large dsDNA viruses, especially Archaeal BJ1 virus (Fig S3), and none of these contigs display a genome size similar to Salterprovirus His2. Hence these contigs are likely to originate from large dsDNA viruses which would have acquired the His2-like aaRS through horizontal gene transfer.

Conversely, several different genes are retrieved for Salterprovirus His1 (Table S5). Three types of contigs can be distinguished (Fig 6) : (i) small circular contigs, (ii) medium-sized circular contigs likely to be complete His1-like genomes and (iii) contigs associated with larger dsDNA viruses. Overall, one gene (His1_gp16) is consistently retrieved in all these contigs, and more generally retrieved in both Senegal and Lake Tyrrell samples. His1_gp16 is a member of the broad AAA+ ATPase family, and similar ATPases are also present in various archaeal plasmids as well as Crenarchaeal viruses (SSV-like fuselloviruses and *Thermococcus prieurii* virus 1). The other His1 genes retrieved are the major capsid gene (His1_gp21), a glycostyltransferase (His1_gp31), and four uncharacterized genes only retrieved in this specific genome so far (His1_gp07, 18, 26 and 32).

Strikingly, all small circular sequences containing a His1_gp16-like gene also harbor a replication-associated gene. Three different types of replication-associated genes are retrieved, each being seemingly associated to a range of genome size (Table S5). For the two smallest type of contigs (between 3 and 4 kb), an archeal plasmid like replication gene can be identified (Table S5), whereas for the largest contigs (4.5 to 5.5 kb), the replication gene identified is similar to *Haloarcula* Pleomorphic virus 1, a small dsDNA archaeal virus (Fig 6A). These different replication-associated genes could thus defined three different types of small circular DNA templates encoding His1_gp16-like gene.

Another group of sequences is formed of circular contigs with a size similar to the genome of His1 and His2 viruses (9 to 15kb ; Table S5 ; Fig 6B). These 4 contigs display at least a gene similar to His1_gp16 and for P6_24 three additional His1-like gene, including His1_gp21, the major capsid protein responsible for the spindle shape of the virion. Interestingly, these contigs all harbor genes similar to phage integrase and archaeal plasmid polymerase.

Finally, a last group of contigs is composed of large sequences containing His1-like genes alongside other Caudovirales and Haloarchaeal plasmid genes (Fig 6C). These genes, when characterized, are associated with genome integration or DNA and protein modification, but no Caudovirales marker gene (capsid, portal or terminase gene) is retrieved in these contigs. Again, one contig (P5_0) harbors a gene similar to the main capsid gene of His1 virus : His1_gp21.

Considering the wide distribution of His1_gp16-like genes, a phylogenetic tree was computed from these sequences to decipher the relationship between these different type of contigs (Fig S4). All the different types of contigs are rather separated on the tree, which seem to indicate that a limited number of gene transfer would have occurred. Surprisingly, sequences similar to His1_gp16 were scarcely retrieved in environmental databases except for a limited number of reads, all from other extreme environments samples (Hydrothermal vents, Saltern viromes and metagenomes ; Table S5). Therefore, it is tempting to speculate that His1_gp16 codes for an ATPase specifically adapted toward an optimal activity in extreme conditions, thus exchanged and conserved through a wide range of genomes from such environments.

Discussion

Environmental viral communities are still largely uncharacterized, and fundamental questions such as the global distribution of viruses or their actual genetic and genomic richness are still open. Unprecedented possibilities to address such questions are now offered by metagenomics approaches associated with high-depth sequencing and bioinformatics meta-analyses, which provided here valuable insights into hyperalophilic viral communities. A global picture of the halophilic viral pan-genome could thus be drawn from the community to the genotype level, forming a useful complement to the isolation and cultivation-based informations available on these communities.

Haloviruses are still largely uncharacterized

Even though massive sequencing allowed us to assemble full length proteins instead of using short reads, most of the predicted ORFs are still unknown. Three non exclusive hypotheses can explain such a high proportion of unknown genes : (i) the lack of close enough references in databases, (ii) the specificity of viruses from these ecosystems and (iii) a high genetic diversity within these communities. Indeed, haloarcheal viruses remain poorly explored, with only 14 haloviral genomes sequenced to date that serve as references. Interestingly, 5 of these individual halophilic viruses recruit as much as 33% of the affiliated ORFs in three high salinity viromes from Senegal. Moreover, completely sequenced haloviruses were previously shown to share only a minority of their genes with other viruses. Both the specificity of haloviruses and the paucity of references for these viruses could explain why a vast majority of hypersaline virome sequences are not similar to any viral proteins of current databases.

If some viral strains typical from hypersaline environments were indeed retrieved in these viromes, some other are absent or quasi absent, like HRPV2 to 6 or virus SH1. This unexpected absence could be linked to a methodological bias in the virome preparation, or to high genomic divergence between viruses infecting even closely related hosts.

Salinity is a major factor driving the composition of viral communities

In this study, viral communities were investigated in two natural sites in Senegal (West Africa) covering a salinity gradient ranging from “low” hypersaline (8%) to near salt saturation (36%). Salinity is known to influence both macro-organisms and microbial community structure (Wang et al., 2011), and was shown to be the main environmental factor influencing micro-organisms communities both in natural hypersaline lake (Wu, Zwart, Schauer, Kamst-van Agterveld, & Hahn, 2006) and artifical hypersaline ponds (Benlloch et al., 2002). Meta-

analyses involving multiple approaches identified two main groups of hypersaline ponds from their micro-organisms communities : low hypersaline environments (4-15%) and high hypersaline ones (22-37%). Based on the specificity of viral-host interactions, corresponding viral communities were thought to exhibit a similar pattern. Indeed, viromes samples from high hypersaline ponds (P5, 6 7 and 8, salinity over 22 %) were more similar between them than with the low salinity sample (P2, 8%). We hypothesize that this trend is mostly due to an adaptation of viruses to the hyperhalophilic host communities, and that the impact of salinity on viruses is thus mostly indirect.

Halophilic viral pan-genome is consistent across time and space

The high percentages of unknowns ORFs observed hinder the comparison of metagenomic samples by traditional methods that map sequencing reads to known, annotated references (Mokili, Rohwer, & Dutilh, 2012). To tackle this bias, we used a reference-independent methodology (Roux et al., 2011) to compare the 6 viromes generated here between them and to a large collection of 35 available aquatic viromes ranging from freshwater to hypersaline. As precluded by Santos *et al.* (Santos et al., 2010), there is a strong similarity in term of genes encoded by viral genomes of the different hypersaline ponds samples around the globe, as observed for freshwater environments (Roux et al., 2012).

In addition to the clear separation of highly saline ecosystems from the rest, a biogeographic pattern seems to exist for the high hypersaline viromes out of the same analysis. Indeed, hypersaline viromes sampled from three distant locations are separated. Furthermore, genetic distances between high salinity viromes are higher than within freshwater, seawater and low-to-medium salinity viromes. The divergence between the Senegal viromes are consistent with the broad salinity range. In contrast, Lake Tyrrell viromes that were collected at the same location but at different dates, appear very closely related. Conversely, Emerson and colleagues in the analysis of these Lake Tyrrell viromes estimated that their viromes presented an important internal diversity (Emerson et al., 2012). These seemingly contradictory results can be explained by the level of genetic similarity studied : in our whole-virome comparison, a BLAST is used to detected similarity between virome sequences at the protein level (tBLASTx), whereas in Emerson *et al.* study, similarities are observed at a nucleotide level, and with very stringent threshold (90 % of nucleotide identity). Thus, halophilic viral communities seem to be stable in term of gene content and functional potential, but rapidly changing at a nucleotide level. This trade-off between global adaptation to environmental parameters and local adaptation to host strains could lead to the pattern observed in the virome comparison.

Halophilic clades and horizontal gene transfer as keys to the halophilic viral pan-genome stability

This genetic homogeneity of halophilic viral communities seems to originate from both the existence of specialized clades and a high rate of lateral gene transfer between unrelated genomes within hypersaline ponds.

Caudovirales diversity gives an examples of specialization within a large viral family. Previous virome studies from seawater or freshwater described the presence of a high number of different Caudovirales strains in single samples (see for example (Angly et al., 2006; López-Bueno et al., 2009; Roux et al., 2012)). In this study, this pattern was only observed for Senegal P2 and Lake Tyrrell samples, two kind of environments which are either constantly (P2) or temporarily but recurrently (Lake Tyrrell) of medium salinity. Conversely, only a limited number of clades of Caudovirales was retrieved in high hypersaline systems. These groups appear to be highly specific to high hypersaline ponds. Moreover, the gene conservation within each haloviral group was very good, and no biogeographic pattern could be observed when comparing Caudovirales genomes fragments from Santa Pola, Lake Tyrrell and Senegal samples.

Contigs associated with Salterprovirus are on the other hand informative toward the extent of lateral gene transfer that can occur within hypersaline environments. An extensive gene exchange between tailed viruses and other types of viruses or genetic elements such as plasmids seem to take place in hypersaline environments. The study of an isolated halophage, Natrialba phage PhiCh1, had already assessed an evolutionary link between this phage genome and several archaeal plasmids (Klein, Baranyi, Greineder, Scholz, & Witte, 2002). The different observations from this study both confirm and extend these observations. Similarly, high level of lateral gene transfer between halophilic micro-organisms was already assumed from the study of complete genomes (Mongodin et al., 2005), and is generally considered as important in the adaptation and survival of micro-organisms in these environments. Our results confirmed that viruses, which are generally considered as major actors of horizontal gene transfer events, could be especially important in high hypersaline environments where they seem to carry, share and transfer genes specific for these extreme conditions. Therefore, these uncharacterized genes conserved between hypersaline ponds around the world, such as gp16-like gene of His1viruses, should be further investigated.

Conclusion

All taken together, the existence of specific genes, specific clades of caudovirales, and the gathering in the comparison based on whole dataset strongly indicates that viral communities of hypersaline environments are highly specific. As noted for marine to freshwater environments, it is likely that transition from low hypersaline to high hypersaline ponds are rare, in so that any non-adapted virus in such extreme conditions as high salinity would probably not be able to multiply. Nevertheless, the homogeneity of hypersaline ponds sampled in different locations geographically separated means that some viral gene exchange must occur at a worldwide level. Thus, even though hypersaline environments have a very ancient geological divergence, hypersaline viral communities could be considered as a genetic continuum between the different hypersaline ponds around the globe, with an exchange rate between

distant ponds likely to be low but not null.

Material & Methods

Sample sites. Samples were taken from 6 locations in Senegal, at the end of the dry season (May 2011) and display salinities ranging from 8 to 36 % (Table 1, Fig S1). The salinity was measured using a hand refractometer. The P2 sample was collected in an inverse tropical estuary (Sine Saloum) and displays a salinity of 8%, *i.e.* about twice seawater salinity. P5, P7, P8 and P9 were collected in individual small marine solar saltern that consist of shallow ponds where seawater is concentrated until sodium chloride is precipitated. The salt concentration of these separated ponds ranges from 26 to 36% the difference being probably due to the size of these ponds. P6 was sampled in Lake Retba, a natural salt lake, with a salinity of 29%. The physico-chemical properties of the different samples are rather uniform, with the main salt components being Na^+ and Cl^- for all samples except for P9, which main cation is Mg^{++}.

DNA extraction and sequencing. Water samples were filtered on 0.2 µm filters followed by PEG precipitation (Colombet et al., 2007). Viral concentrates were then re-filtrated on a 0.2 µm screen, to remove any remaining cellular micro-organisms, then quantified by flow cytometry (Brussaard, 2004). The number of Virus-like particles (VLP) ranged from 9.10^{-6} to 1.10^{-9} per mL. Viral concentrates were treated with DNAseI (Invitrogen) to remove external DNA fragments. Encapsidated DNA was freed via thermal shock then purified using a QuiAmp DNA mini kit (Quiagen). Samples were then treated with RNAse and dialysed on ester cellulose filter 0.025 µm (Millipore). DNA amplification was run with a GenomiPhi Kit (GE Healthcare) which produced non-specific amplification through polymerase phi29. The six DNA preparations underwent library construction and sequencing at GATC Biotech (Germany) on a single lane of Illumina HiSeq2000.

Data sets and assembly. The amount of DNA sequenced for the six P sample was quite homogeneous, the number of 100-bp paired-end reads ranging from 44 to 62 millions. These reads were checked for quality with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/) and low-quality right ends were removed with the FastX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html ; threshold of 30). Each sample was then independently assembled using Idba_ud (Peng, Leung, Yiu, & Chin, 2012). Even with comparable number of reads, the number of contigs assembled are drastically different, ranging from 17,173 for P9 to 89,754 for P2 (Table 1). Unexpectedly, this lower number of contigs assembled is not associated with an increase in the contigs average size, meaning that despite the higher redundancy of the DNA pool sequenced, the assembly of long genomic fragment is still difficult. Nevertheless, large genomic fragments are generated for each sample, as indicated by the maximum contig size spanning from 38 Kb for P6 to 60 Kp for P2. ORFs were predicted for each contig using MetaGeneAnnotator (Noguchi, Taniguchi, & Itoh, 2008).

Taxonomic affiliation and contig filtering. Each predicted ORF was compared to the nr database and to the refseq viral genome proteins database (both downloaded from the NCBI) via BLASTp, with a threshold of 50 on bitscore and 0.001 on e-value. A small but consistent ratio of contigs was found to be clearly associated with contamination by cellular genomic fragments, such that a contig filtering was processed. Only contig for which more than three genes are predicted, similar to at least a viral genome or with less than half of predicted genes affiliated to a cellular genome were selected. This first automatic filtering was manually curated to keep all contigs likely to come from viral genomes, and remove all clear contamination. All results presented in the manuscript are based on this filtered set of contigs.

All filtered contigs were uploaded to the Metavir server ((Roux et al., 2011) ; http://metavir-meb.univ-bpclermont.fr ; Project Archevir ; note that only contigs greater than 500 bp are kept in this process). Reads from the Lake Tyrrell samples (Emerson et al., 2012) were downloaded from SRA, assembled with Newbler (454 Roche Software) or Idba_ud (Peng, Leung, Yiu, & Chin, 2010) for 454 or Hi-Seq sequenced datasets respectively, and similarly uploaded to Metavir (Project Lake Tyrrell). Santa Pola fosmids (Garcia-Heredia et al., 2012) were already publicly available on the same web-server (Project "Others", virome "Saltern ponds fosmids").

Whole-virome comparison. Sub-samples of 50,000 reads of 100bp were selected and used for tBLASTx virome comparison on the Metavir web-server (Roux et al., 2011). Briefly, tBLASTx pairwise comparison were computed, and a mean score was deduced from each of these comparison. Such comparison is especially fitted for viromes as it takes into account both identified and unidentified sequences (which can represent the major part of a viral metagenome). The matrix of these scores was then used in a MDS to plot viromes (metaMDS function with a Bray-Curtis dissimilarity index of the package vegan from R software ; (Oksanen et al., 2008)).We found a very good linear correlation between the distances on the plot and the dissimilarities in the matrix for each pair of point (r^2=0.952, Fig S5), confirming that the plot was a correct representation of the genetic distance between the different datasets. Correlation with salinity was assessed with function envfit from the same package (vector coordinates for salinity : 0.937 for axis1, 0.350 for axis2, 999 permutations).

Phylogenetic trees and contigs analysis. All complete or partial genomes of viruses from hypersaline ecosystems were considered in this analysis. In addition to contigs from the present viromes and from the Lake Tyrrell viromes (Australia; (Emerson et al., 2012)), 42 fosmid sequences of 40 Kb obtained from Santa Pola pond (Spain; (Garcia-Heredia et al., 2012)) were also used in the analysis. Using phylogenetic analysis of marker genes and comparison of genomic maps available on Metavir (Roux et al., 2011), the two major groups of viruses present in the large contigs of the Senegal's viromes were then thoroughly analyzed : head-tail dsDNA viruses (mostly affiliated to *Caudovirales)* and Salterproviruses.

Both phylogenetic trees (TerL and gp16) were computed with FastTree2 (Price, Dehal, & Arkin, 2010) from a multiple alignement computed with Muscle (Edgar, 2004) and manually curated. All viromes sequences detected as similar to His1_gp16 were use as query in a BLASTp against NCBI Env_NR database and a tBLASTn against NCBI Env_nt through the Camera Web Portal (Sun et al., 2011). A threshold of 10^{-3} on e-value was used, and no significantly similar sequences were retrieved in Env_NR.

Acknowledgments

The authors gratefully acknowledge the IRD for their help during the sampling campaign. This study was funded by a program grant (Archevir) from the French "Fondation pour la Recherche sur la Biodiversité" (FRB). SR was supported by a PhD grant from the French defense procurement agency (DGA, Direction Générale de l'Armement). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Aalto, A. P., Bitto, D., Ravantti, J. J., Bamford, D. H., Huiskonen, J. T., & Oksanen, H. M. (2012). Snapshot of virus evolution in hypersaline environments from the characterization of a membrane-containing Salisaeta icosahedral phage 1. Proceedings of the National Academy of Sciences of the United States of America, 109(18), 7079–84. doi:10.1073/pnas.1120174109

Angly, F. E., Felts, B., Breitbart, M., Salamon, P., Edwards, R. A., Carlson, C., Chan, A. M., et al. (2006). The marine viromes of four oceanic regions. PLoS biology, 4(11), e368.

Arslan, D., Legendre, M., Seltzer, V., Abergel, C., & Claverie, J.-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. Proceedings of the National Academy of Sciences of the United States of America, 108(42), 17486–91. doi:10.1073/pnas.1110889108

Bath, C., Cukalac, T., Porter, K., & Dyall-Smith, M. L. (2006). His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. Virology, 350(1), 228–39. doi:10.1016/j.virol.2006.02.005

Baxter, B. K., Mangalea, M. R., Willcox, S., Sabet, S., Nagoulat, M.-N., & Griffith, J. D. (2011). Haloviruses of Great Salt Lake: A Model for Understanding Viral Diversity. In A. Ventosa, A. Oren, & Y. Ma (Eds.), Halophiles and Hypersaline Environments (Springer., pp. 173–190). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-642-20198-1

Benlloch, S., López-López, A., Casamayor, E. O., Øvreås, L., Goddard, V., Daae, F. L., Smerdon, G., et al. (2002). Prokaryotic genetic diversity throughout the salinity gradient of a coastal solar saltern. Environmental microbiology, 4(6), 349–60.

Bettarel, Y., Bouvier, T., Bouvier, C., Carré, C., Desnues, A., Domaizon, I., Jacquet, S., et al. (2011). Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. FEMS microbiology ecology, 76(2), 360–72. doi:10.1111/j.1574-6941.2011.01054.x

Bolhuis, H., Poele, E. M. Te, & Rodriguez-Valera, F. (2004). Isolation and cultivation of Walsby's square archaeon. Environmental microbiology, 6(12), 1287–91. doi:10.1111/j.1462-2920.2004.00692.x

Boujelben, I., Yarza, P., Almansa, C., Villamor, J., Maalej, S., Antón, J., & Santos, F. (2012). Virioplankton community structure in Tunisian solar salterns. Applied and environmental microbiology, 78(20), 7429–37. doi:10.1128/AEM.01793-12

Brussaard, C. P. D. (2004). Optimization of procedures for counting viruses by flow cytometry. Applied and environmental microbiology, 70(3), 1506–1513.

Casamayor, E. O., Massana, R., Benlloch, S., Øvreås, L., Díez, B., Goddard, V. J., Gasol, J. M., et al. (2002). Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. Environmental microbiology, 4(6), 338–48.

Cho, B. C. (2005). Heterotrophic Flagellates in Hypersaline Waters. In N. Gunde-Cimerman, A. Oren, & A. Plemenitaš (Eds.), Adaptation to Life at High Salt Concentrations in Archaea, Bacteria, and Eukarya (Springer., pp. 541–549). Springer Netherlands.

Colombet, J., Robin, A., Lavie, L., Bettarel, Y., Cauchie, H. M., & Sime-Ngando, T. (2007). Virioplankton "pegylation": use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. Journal of microbiological methods, 71(3), 212–219.

Edgar, R. C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5, 113.

Emerson, J. B., Thomas, B. C., Andrade, K., Allen, E. E., Heidelberg, K. B., & Banfield, J. F. (2012). Metagenomic assembly reveals dynamic viral populations in hypersaline systems. Applied and environmental microbiology, 78(17), 6309 – 6320. doi:10.1128/AEM.01212-12

Fischer, M. G., Allen, M. J., Wilson, W. H., & Suttle, C. A. (2010). Giant virus with a remarkable complement of genes infects marine zooplankton, 107(45), 1–6. doi:10.1073/pnas.1007615107/-/DCSupplemental.www.pnas.org/cgi/doi/10.1073/pnas.1007615107

Garcia-Heredia, I., Martin-Cuadrado, A.-B., Mojica, F. J. M., Santos, F., Mira, A., Antón, J., & Rodriguez-Valera, F. (2012). Reconstructing Viral Genomes from the Environment Using Fosmid Clones: The Case of Haloviruses. (M. R. Liles, Ed.)PLoS ONE, 7(3), e33802. doi:10.1371/journal.pone.0033802

Hendrix, R. W. (2009). Jumbo bacteriophages. Current topics in microbiology and immunology, 328, 229–40.

Klein, R., Baranyi, U., Greineder, B., Scholz, H., & Witte, A. (2002). Natrialba magadii virus PhiCh1 : first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. Molecular Microbiology, 45(3), 851–863.

Laybourn-Parry, J., Hofer, J. S., & Sommaruga, R. (2001). Viruses in the plankton of freshwater and saline Antarctic lakes. Freshwater Biologyre, 46, 1279–1287.

López-Bueno, A., Tamames, J., Velázquez, D., Moya, A., Quesada, A., & Alcamí, A. (2009). High diversity of the viral community from an Antarctic lake. Science, 326(5954), 858–61. doi:10.1126/science.1179287

Mokili, J. L., Rohwer, F., & Dutilh, B. E. (2012). Metagenomics and future perspectives in virus discovery. Current Opinion in Virology, 2(1), 63–77. doi:10.1016/j.coviro.2011.12.004

Mongodin, E. F., Nelson, K. E., Daugherty, S., Deboy, R. T., Wister, J., Khouri, H., Weidman, J., et al. (2005). The genome of Salinibacter ruber: convergence and gene exchange among hyperhalophilic bacteria and archaea. Proceedings of the National Academy of Sciences of the United States of America, 102(50), 18147–52. doi:10.1073/pnas.0509073102

Noguchi, H., Taniguchi, T., & Itoh, T. (2008). MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA research, 15(6), 387–96. doi:10.1093/dnares/dsn027

Oh, D., Porter, K., Russ, B., Burns, D., & Dyall-Smith, M. (2010). Diversity of Haloquadratum and other haloarchaea in three, geographically distant, Australian saltern crystallizer ponds. Extremophiles : life under extreme conditions, 14(2), 161–9. doi:10.1007/s00792-009-0295-6

Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Simpson, G. L., Solymos, P., Stevens, M. H. H., et al. (2008). The vegan Package.

Oren, A. (2008). Microbial life at high salt concentrations: phylogenetic and metabolic diversity. Saline systems, 4, 2. doi:10.1186/1746-1448-4-2

Oren, A., Sørensen, K. B., Canfield, D. E., Teske, A. P., Ionescu, D., Lipski, A., & Altendorf, K. (2009). Microbial communities and processes within a hypersaline gypsum crust in a saltern evaporation pond (Eilat, Israel). Hydrobiologia, 626(1), 15–26. doi:10.1007/s10750-009-9734-8

Pagaling, E., Haigh, R. D., Grant, W. D., Cowan, D. a, Jones, B. E., Ma, Y., Ventosa, A., et al. (2007). Sequence analysis of an Archaeal virus isolated from a hypersaline lake in Inner Mongolia, China. BMC genomics, 8, 410. doi:10.1186/1471-2164-8-410

Pedrós-Alió, C., Calderón-Paz, J., MacLean, M., Medina, G., Marrasé, C., Gasol, J., & Guixa-Boixereu, N. (2000). The microbial food web along salinity gradients. FEMS microbiology ecology, 32(2), 143–155.

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2010). IDBA – A Practical Iterative de Bruijn Graph De Novo Assembler. RECOMB (pp. 426–440).

Peng, Y., Leung, H. C. M., Yiu, S. M., & Chin, F. Y. L. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics, 28(11), 1420–1428.

Pietilä, M. K., Laurinmäki, P., Russell, D. a, Ko, C.-C., Jacobs-Sera, D., Butcher, S. J., Bamford, D. H., et al. (2013). Insights into head-tailed viruses infecting extremely halophilic archaea. Journal of virology, 87(6), 3248–60. doi:10.1128/JVI.03397-12

Pina, M., Bize, A., Forterre, P., & Prangishvili, D. (2011). The archeoviruses. FEMS microbiology reviews, 35(6), 1035–54. doi:10.1111/j.1574-6976.2011.00280.x

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One, 5(3), e9490.

Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., et al. (2004). The 1.2-megabase genome sequence of Mimivirus. Science, 306(5700), 1344–50. doi:10.1126/science.1101485

Rodriguez-Brito, B., Li, L., Wegley, L., Furlan, M., Angly, F., Breitbart, M., Buchanan, J., et al. (2010). Viral and microbial community dynamics in four aquatic environments. The ISME journal, 4(6), 739–51. doi:10.1038/ismej.2010.1

Rohwer, F., Prangishvili, D., & Lindell, D. (2009). Roles of viruses in the environment. Environmental microbiology, 11(11), 2771–4. doi:10.1111/j.1462-2920.2009.02101.x

Roine, E., Kukkaro, P., Paulin, L., Laurinavicius, S., Domanska, A., Somerharju, P., & Bamford, D. H. (2010). New, closely related haloarchaeal viral elements with different nucleic Acid types. Journal of virology, 84(7), 3682–9. doi:10.1128/JVI.01879-09

Roine, E., & Oksanen, H. M. (2011). Halophiles and Hypersaline Environments. (A. Ventosa, A. Oren, & Y. Ma, Eds.), 56(Viikinkaari 5), 153–172. doi:10.1007/978-3-642-20198-1

Roux, S., Enault, F., Robin, A., Ravet, V., Personnic, S., Theil, S., Colombet, J., et al. (2012). Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. PLoS One, 7(3), e33641.

Roux, S., Faubladier, M., Mahul, A., Paulhe, N., Bernard, A., Debroas, D., & Enault, F. (2011). Metavir: a web server dedicated to virome analysis. Bioinformatics, 27(21), 3074–3075. doi:10.1093/bioinformatics/btr519

Santos, F., Yarza, P., Parro, V., Briones, C., & Antón, J. (2010). The metavirome of a hypersaline environment. Environmental microbiology, 12(11), 2965–76. doi:10.1111/j.1462-2920.2010.02273.x

Schapira, M., Buscot, M., Leterme, S., Pollet, T., Chapperon, C., & Seuront, L. (2008). Distribution of heterotrophic bacteria and virus-like particles along a salinity gradient in a hypersaline coastal lagoon. Aquatic Microbial Ecology, 54(February), 171–183. doi:10.3354/ame01262

Sime-Ngando, T., Lucas, S., Robin, A., Tucker, K. P., Colombet, J., Bettarel, Y., Desmond, E., et al. (2010). Diversity of virus-host systems in hypersaline Lake Retba, Senegal. Environmental microbiology, 13(8), 1956–1972.

Sun, S., Chen, J., Li, W., Altintas, I., Lin, A., Peltier, S., Stocks, K., et al. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. Nucleic acids research, 39(Database issue), D546–51. doi:10.1093/nar/gkq1102

Tang, S., Nuttall, S., & Dyall-smith, M. (2004). Haloviruses HF1 and HF2: Evidence for a Recent and Large Recombination Event. Journal of bacteriology, 186(9), 2810–2817. doi:10.1128/JB.186.9.2810

Wang, J., Yang, D., Zhang, Y., Shen, J., Van der Gast, C., Hahn, M. W., & Wu, Q. (2011). Do patterns of bacterial diversity along salinity gradients differ from those observed for macroorganisms? PloS one, 6(11), e27597. doi:10.1371/journal.pone.0027597

Wu, Q. L., Zwart, G., Schauer, M., Kamst-van Agterveld, M. P., & Hahn, M. W. (2006). Bacterioplankton community composition along a salinity gradient of sixteen high-mountain lakes located on the Tibetan Plateau, China. Applied and environmental microbiology, 72(8), 5478–85. doi:10.1128/AEM.00767-06

Adaptation génomique des virus aux milieux hypersalins

L'analyse des viromes du programme Archevir associée aux résultats précédents obtenus par approche de métagénomique sur différents bassins hypersalins (Rodriguez-Brito *et al.*, 2010; Emerson *et al.*, 2012; Garcia-Heredia *et al.*, 2012) permet de dégager de grandes tendances générales concernant ces communautés virales. Ces dernières semblent spécifiques, et donc très adaptées à ce type d'environnement, avec une certaine homogénéité autour du globe. Si une part importante de ces virus hypersalins est encore non identifiée, un certain nombre de clades ou groupes viraux typiques de ces milieux (et uniquement détectés dans ces échantillons hypersalins) ont pu être mis en évidence, comme par exemple les *Salterprovirus,* ou encore différents clades au sein des *Caudovirales.*

La deuxième grande tendance dégagée de l'analyse de ces différents viromes et des différents génomes complets de virus hypersalins (Klein *et al.*, 2002; Bath *et al.*, 2006) est le fort taux de transfert horizontal entre virus de types différents (différentes natures et tailles de génomes, ou différents types de capsides), mais aussi entre génomes viraux et plasmides et génomes de bactéries et d'archées. Dans ce cadre, les études de viromes ont permis d'identifier plusieurs nouveaux types de virus hypersalins, mais aussi une série de gènes visiblement spécifiques de ces milieux hypersalins conservés et transférés entre différents types de génomes. Des études complémentaires de ces différents gènes seraient ainsi nécessaires pour déterminer leur fonction, potentiellement intéressante d'un point de vue biotechnologique de par l'adaptation de ces gènes à ces milieux extrêmes.

Distribution globale et adaptations locales des communautés virales aquatiques

De manière générale, une comparaison de séquences des viromes aquatiques permet de mettre en évidence une séparation de ces communautés en fonction de la salinité de l'échantillon, et non en fonction de la localisation de ces échantillons (Figure III.2). Il semble toutefois exister au sein des viromes fortement hypersalins et d'eau douce des points extrêmes (comme P9 au sein d'Archevir, ou l'échantillon “printemps” du lac Limnopolar), pour lesquels le pan-génome viral apparaît unique. Ces points correspondent aux conditions environnementales les plus singulières, puisque le point P9 présentait une salinité très forte et une composition physico-chimique spécifique, tandis que l'échantillon “printemps” du lac Limnopolar correspond à une période au cours de laquelle le lac est recouvert de glace. À

l'inverse, les viromes marins semblent très homogènes, et uniquement séparés entre échantillons de surface et profonds, et les viromes issus du lac Tyrrell, bien que recouvrant trois ans de prélèvements, sont eux aussi très proches. Ainsi, les conditions environnementales semblent être le facteur principal de différenciation des communautés virales, plus que la distance (physique ou temporelle) entre les prélèvements.

Il est à noter que cette répartition des différents viromes en fonction de la salinité intervient alors même que ces différents jeux de données ont pour la plupart été préparés par des laboratoires différents avec des protocoles différents, que ce soit au niveau de la porosité des filtres utilisés, des méthodes de précipitation des capsides virales, ou encore des méthodes d'amplification et de séquençage utilisées. Ainsi, s'il faut se garder de toute sur-interprétation, l'observation malgré cet ensemble de biais potentiels d'un regroupement clair et constant des viromes issus du même type de biome indique que cette similarité des pangénomes viraux entre échantillons de même niveau et type de salinité est véritablement une tendance forte et importante.

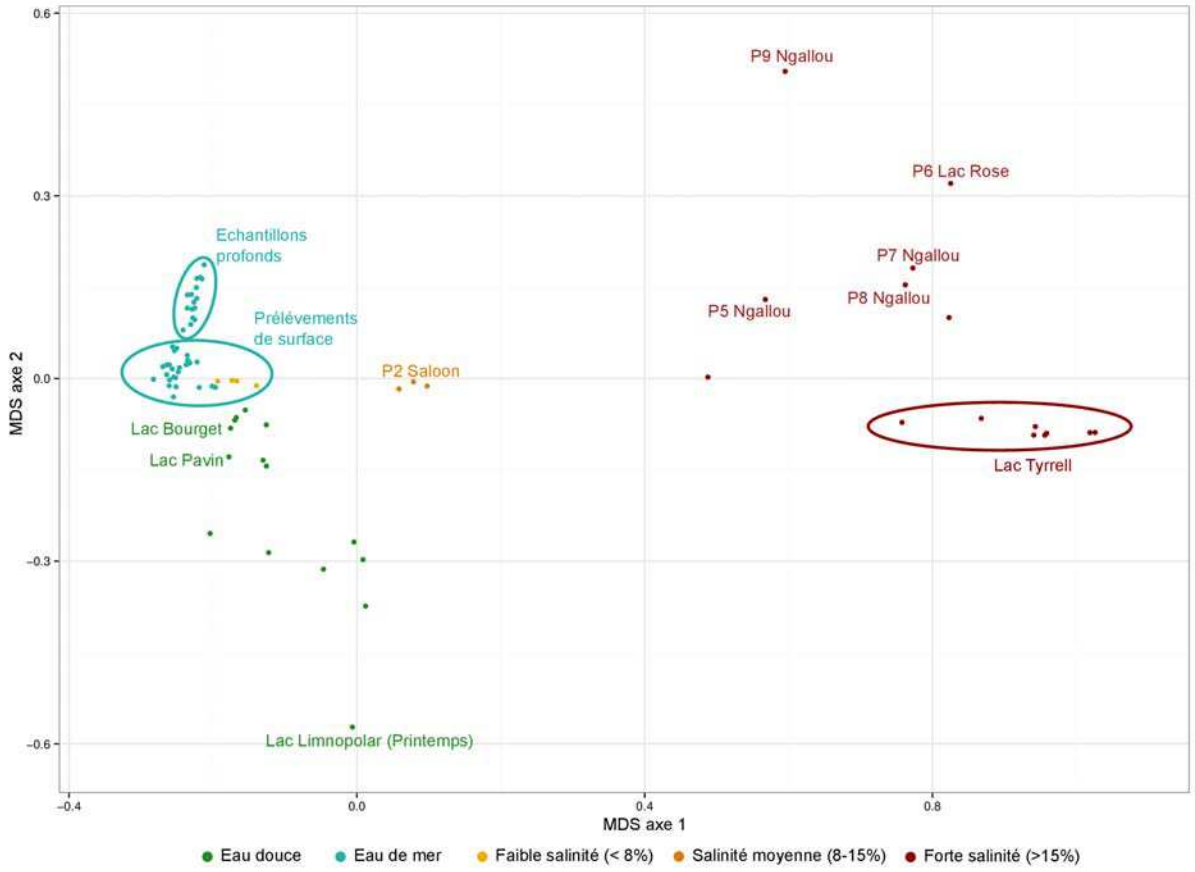


Figure III.2 : Comparaison globale de viromes aquatiques. *Les séquences des viromes sont comparées deux à deux via un tBLASTx de sous-échantillons normalisés (50 000 séquences de 100 pb). Un MDS (positionnement multidimensionnel) est ensuite réalisé, qui permet de positionner les viromes sur un graphique en deux dimensions au sein duquel les distances entre points sont le plus proche possible des distances génétiques entre viromes. Les études présentées dans les parties précédentes sont identifiées avec le nom des lacs (Bourget et Pavin) et le nom des six échantillons hypersalins (P2 à P9).*

Sélection des communautés virales par les communautés d'hôtes

Cette absence de distribution biogéographique pour les communautés virales, du moins à ce niveau d'analyse relativement large (similarité de séquences protéiques), semble ainsi confirmer que de par le nombre extrêmement important de virions, leur petite taille et leur poids extrêmement réduit, les virus sont susceptibles de se disperser à une échelle globale. Si les virus sont effectivement ubiquistes, le principal facteur pouvant expliquer la différenciation des communautés virales est alors la composition des communautés d'hôtes potentiels. Ainsi, il paraît raisonnable d'envisager que la salinité du milieu soit un facteur déterminant pour la composition de la communauté virale d'un milieu aquatique de par l'influence de ce paramètre sur les communautés d'hôtes, notamment de micro-organismes (Casamayor *et al.*, 2002; Wang *et al.*, 2011). La formule de Beijerinck et Becking "tout est partout, mais l'environnement sélectionne" pourrait être adaptée aux virus et simplifiée sous la forme "tous les virus sont partout, mais sélectionnés par les communautés d'hôtes". Cette idée d'une distribution globale des virus, dont les communautés locales seraient ensuite principalement modelées par l'adaptation aux communautés d'hôtes présentes dans le milieu, a été précédemment évoquée (Breitbart & Rohwer, 2005; Vega Thurber, 2009), et les observations réalisées lors de l'étude de groupes spécifiques de virus semblent correspondre à ce schéma de distribution (Breitbart *et al.*, 2004b; Short & Suttle, 2005; Snyder *et al.*, 2007).

Ces résultats peuvent être complétés par les analyses d'infection croisées réalisées entre virus et hôtes issus de différents prélèvements environnementaux. Ainsi, en 2011, Atasanova et collaborateurs ont analysé le potentiel infectieux de phages de milieux hypersalins prélevés tout autour du globe, en utilisant des hôtes de ces mêmes milieux (Atanasova *et al.*, 2012). Les auteurs ont ainsi pu montrer qu'un virus d'un bassin hypersalin pouvait la plupart du temps infecter un hôte d'un autre bassin, quelle que soit la distance entre ces prélèvements. Ces résultats confirment ainsi les analyses métagénomiques, et semblent indiquer que les bassins hypersalins autour du monde, s'ils sont géographiquement séparés, forment un habitat unique et un continuum génétique et évolutif, au moins pour ce qui est des communautés virales.

En 2004, Sano et collaborateurs avaient également réalisés des tests d'infectivité à partir de phages issus d'échantillons environnementaux, mais cette fois entre différents types d'écosystèmes (Sano *et al.*, 2004). De manière surprenante, la plupart des virus issus d'un biome étaient capable d'infecter les membres d'une communauté d'hôte issue d'un écosystème différent. Ainsi, les différences entre communautés virales de biomes différents pourraient ne pas être liées à une incapacité pour les virus d'un type d'environnement d'infecter un hôte d'un autre écosystème, mais plutôt associées à une adaptation de ces communautés virales à

chaque milieu et communauté d'hôtes, et donc une sélection par compétition de ces différents virus. Ainsi, la course à l'armement régulièrement décrite entre virus et hôte existe également certainement entre les virus eux-même. Cet aspect de compétition entre virus est ainsi certainement à prendre en compte lors des analyses de spécificité d'hôte et de capacité d'infection afin d'appréhender au mieux les liens entre populations d'hôtes et populations virales à l'échelle des communautés.

Distribution des virus au sein d'un biome : l'exemple du milieu marin

Si les communautés virales semblent principalement structurées par type d'environnement au niveau global, l'influence de la distance géographique sur la distribution de populations virales pourrait tout de même exister au sein d'un type de milieu. L'étude récente de la diversité bactérienne, archéenne et virale par l'intermédiaire de profils génétiques (T-RFLP et RAPD-PCR) sur 13 sites couvrant 3 200km au nord-est du Canada (de l'Océan Atlantique Nord à l'Océan Arctique) a ainsi mis en avant une distribution des virus qui semblait liée à la distance entre les points plutôt qu'aux paramètres environnementaux, alors que la distance ne semblait pas du tout impacter la distribution des bactéries et archées (Winter *et al.*, 2013). De même, il a été décrit une distribution “biogéographique” de certains phages, que ce soit à partir de métagénomes issus d'un seul type de prélèvements (Williamson *et al.*, 2008b), ou d'études ciblées sur des couples virus-hôte spécifiques isolés dans différents bassins séparés (Kunin *et al.*, 2008; Held & Whitaker, 2009).

Grâce aux viromes les plus récents (Williamson *et al.*, 2012; Hurwitz & Sullivan, 2013), le même type d'étude est maintenant possible au niveau du pan-génome viral, incorporant uniquement des échantillons appartenant au même grand type d'écosystème (à nouveau le milieu marin). Cet ensemble de viromes comprend des viromes prélevés au sein des océans Pacifique et Indien, de la surface à plusieurs kilomètres de profondeur, et des côtes à la pleine mer, préparés selon le même protocole, et pour lesquels un ensemble de métadonnées est disponible.

La comparaison de ces différents viromes, réalisée à partir de résultats obtenus *via* le serveur Metavir, permet d'illustrer un peu plus précisément les liens entre ces différentes communautés virales (Figure III.3). Très clairement, la distinction principale semble se faire entre les viromes issus de la zone photique et ceux issus de la zone aphotique, avant une séparation au sein des viromes de la zone photique entre les différentes zones d’échantillonnage. Les distances génétiques entre viromes ne sont que très faiblement corrélées aux distance géographiques entre points d’échantillonnage (test de Mantel : corrélation : 0,143 ; *p-value* : 0,044 pour le test de 9 999 permutations), confirmant que la

distance entre deux points n'est pas le principal paramètre influençant la distribution des communautés virales. Une projection des différents paramètres environnementaux peut toutefois être réalisée (Figure III.3).

La distinction entre zones photique et aphotique se retrouve de manière prévisible au niveau des métadonnées, avec une zone aphotique plus profonde, et une température et concentration d'oxygène plus élevée dans la zone photique. De plus, si la longitude et la latitude semblent corrélées avec le deuxième axe de la projection, ces paramètres sont également négativement corrélés avec la température, et ces deux facteurs (localisation et température de l'eau) sont impossibles à dissocier dans ce jeu de données. Il est toutefois intéressant de noter que les viromes issus de prélèvements de surface semblent plus distants entre eux que les viromes de profondeurs. Cette observation doit être nuancée par le fait que le jeu de données comprend plus de viromes de surface issus de prélèvements différents, mais cette tendance à une plus grande diversité au sein de la zone photique semble tout de même exister, notamment lorsque l'on compare les transects LineP et MBARI pour lesquels les deux types d'échantillons sont disponibles.

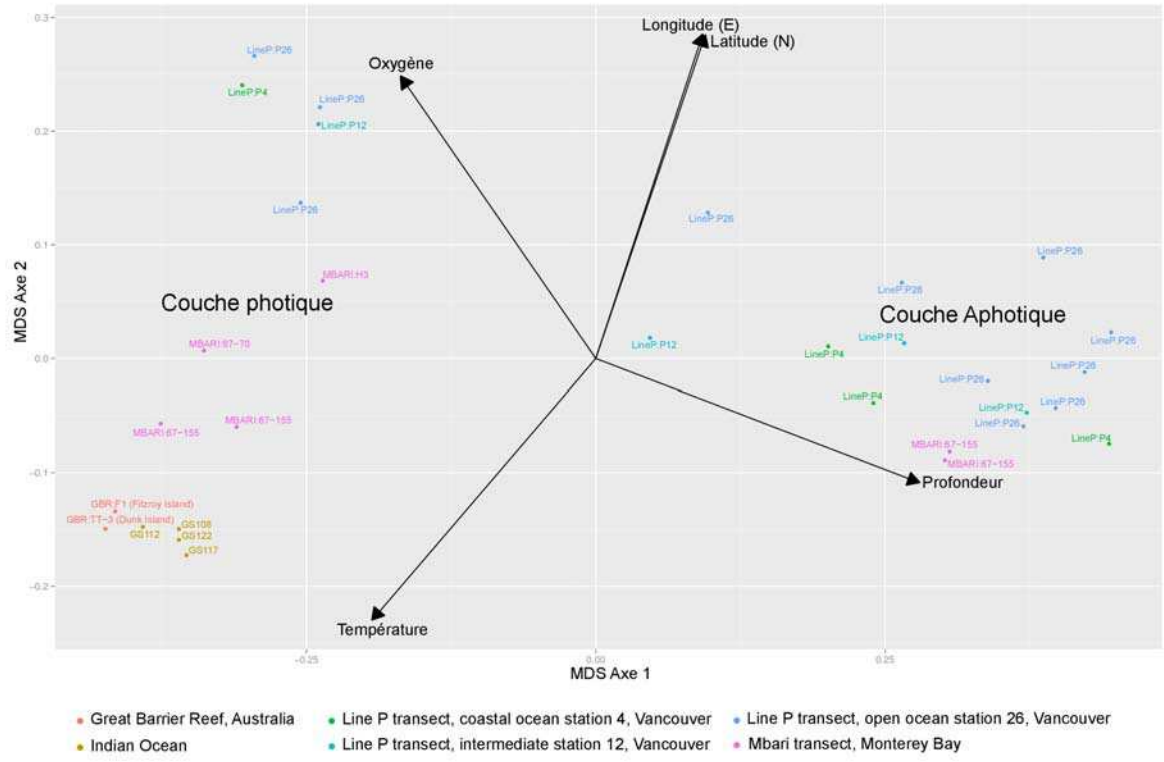


Figure III.3 : Comparaison globale de viromes marins. *Les viromes ont été comparés selon la méthodologie précédemment décrite (tBLASTx de sous-échantillons normalisés, puis NMDS). Les paramètres environnementaux sont ensuite positionnés sur le graphique soit sous la forme d'un vecteur pour les facteurs numériques, soit sous la forme de centre de gravité pour les données qualitatives. Seuls les paramètres dont la corrélation était significative avec la représentation (p-value < 0.001) sont représentés.*

Il est tentant d'associer cette observation à différentes études concernant les cycles viraux dans les océans, qui ont permis d'associer les zones les plus productives plutôt aux cycles lytiques, et les zones où les abondances bactériennes sont plus faibles aux cycles lysogéniques et chroniques (Weinbauer *et al.*, 2003; Williamson *et al.*, 2008a; Payet & Suttle, 2013). Ainsi, la population de virus en surface serait soumis à un turnover plus fréquent, avec une coévolution virus-hôte plus importante du fait des multiples infections, menant rapidement à une spécificité des communautés virales dans chaque zone, que le transfert de virions par les masses d'eau ou par l'atmosphère n'a pas le temps d'homogénéiser. À l'inverse, les communautés benthiques de par leur cycle lysogénique ou chronique, présenteraient globalement plus d'homogénéité à la fois entre les points d'échantillonnage et au sein de la colonne d'eau, puisque les échantillons prélevés à 500, 1 000 et 2 000m sont très proches génétiquement.

La métagénomique permet donc une approche inédite de la diversité génétique et de la distribution des communautés virales aquatiques autour du globe. Ces dernières seraient notamment spécifiques de chaque type d'environnement, et peu impactées par la distance géographique entre deux milieux. Cette spécificité est vraisemblablement liée à une adaptation des communauté virales aux caractéristiques des communautés d'hôtes, associée à une forte compétition entre virus. Au sein de chaque type d'environnement, un équilibre existe certainement entre la capacité de dispersion globale des particules virales et un mouvement de diversification locale lié notamment aux différentes interactions entre hôtes et virus, dont la vitesse sera dépendante du type de cycle infectieux réalisé.

Chapitre IV – Virus à ADN simple brin : diversité et mécanismes d'évolution

L'application des approches de métagénomique à différents types d'échantillons et de milieux a ainsi permis une caractérisation plus précise des communautés virales environnementales. Dans ce cadre, deux groupes principaux sont régulièrement détectés au sein des virus à ADN : les *Caudovirales*, et les petits virus à ADN simple brin. Si la présence de bactériophages à queue était attendue, celle, parfois très importante voire majoritaire, des petits virus à ADN simple brin dans des viromes de différents écosystèmes fut plus surprenante (Angly *et al.*, 2006; Djikeng *et al.*, 2009; López-Bueno *et al.*, 2009; Rosario *et al.*, 2009a; Blinkova *et al.*, 2010; Kim *et al.*, 2011; Ng *et al.*, 2012; Roux *et al.*, 2012). En effet, leur petite taille (virions entre 15 et 30 nm) les rendant difficilement détectables par les méthodes classiques d'étude des communautés environnementales (microscopie électronique et cytométrie en flux), l'abondance de ces virus était jusqu'alors sous-estimée et ces virus étaient considérés comme des acteurs secondaires des écosystèmes.

Les seules connaissances concernant ces petits virus avaient jusqu'alors été obtenues à partir de virus cultivables, d'intérêt économique et clinique, assez peu représentatifs des communautés observées dans les milieux naturels. Ces virus représentent de plus un extrême dans le spectre de taille biologique, puisqu'ils incluent les plus petits pathogènes eucaryotes dont le génome ne code que pour une capside et une protéine de réplication. Ce groupe de virus comprend également les plus petits phages bactériens (familles des *Microviridae* et *Inoviridae*), dont les génomes contiennent moins d'une dizaine de gènes. En mettant en lumière la répartition ubiquiste de ces virus, la métagénomique virale a amené les virologistes à reconsidérer leur importance écologique et suggère que ces virus à la structure génomique minimale pourraient être prévalents dans l'environnement.

Ainsi, nous avons cherché à mieux caractériser ces virus, estimer leur diversité et leur distribution, comprendre quels pouvaient être leurs hôtes dans les milieux aquatiques, et enfin identifier les différents mécanismes concourant à la trajectoire évolutive de leur génome.

Les Microviridae : une famille de phages à ADN simple brin

Les *Microviridae* constituent l'une des deux familles de phages bactériens possédant un génome à ADN simple brin (avec les *Inoviridae*, avec lesquels ils ne partagent pas d'autres caractéristiques). Ils sont formés d'un virion de type icosaédrique d'environ 30 nm de diamètre, renfermant un génome de 5 kb en moyenne (Figure IV.1). La description des premiers Microvirus remonte aux débuts de la virologie, puisque les premières observations du virus emblématique de cette famille, l'Enterobacteria phage phiX174, ont été réalisées à

Paris au cours des années 1920 (Cherwa & Fane, 2011). Ce virus fut déterminant dans l'histoire de la virologie en étant le premier phage doté d'une capside sans queue, mais aussi le premier dont le génome était à ADN simple brin, à une époque où la molécule d'ADN était considérée comme uniquement double brin. Le séquençage de ce génome a également révélé pour la première fois qu'un gène pouvait être "situé au sein d'un autre gène" (Cherwa & Fane, 2011). Par la suite, les Microvirus constitueront des entités modèles pour de nombreuses études sur l'adaptation des génomes viraux (Rokyta *et al.*, 2009), le transfert horizontal de gènes entre souches proches (Rokyta *et al.*, 2006), les mécanismes d'assemblage de la capside (Bernal *et al.*, 2003), l'organisation des gènes au sein du génome (Jaschke *et al.*, 2012), ou encore la résistance des capsides virales dans l'environnement (Lee & Sobsey, 2011).

Paradoxalement, malgré ces nombreuses études portant sur cette famille, la question de la diversité des *Microviridae* a peu été abordée. En 2002, Brentlinger et collaborateurs font ainsi le constat d'une "famille divisée", entre les Microvirus, isolés sur *E. coli* et sujets de nombreuses expériences, et un petit groupe de virus plus lointains, infectant des bactéries

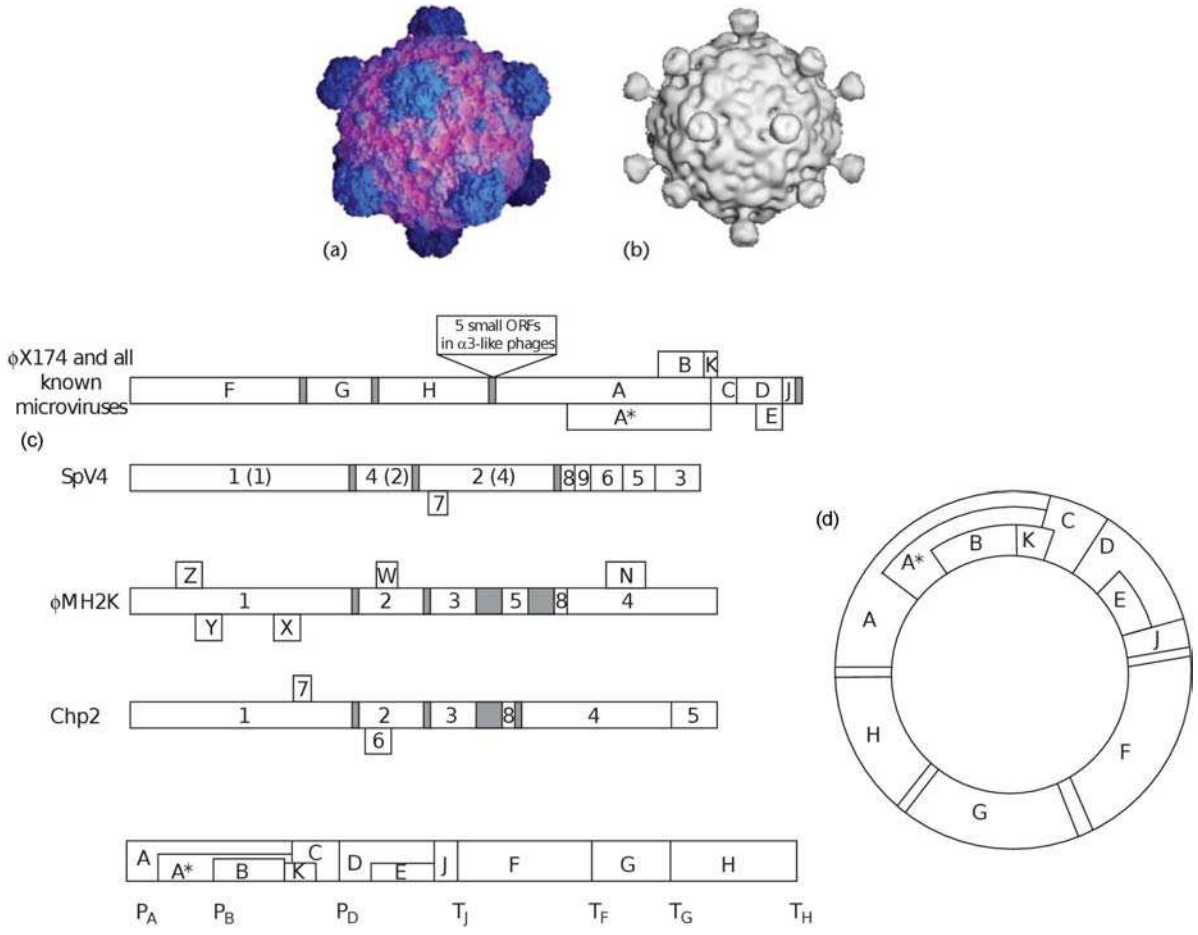


Figure IV.1 : Structure du virion et génome des Microviridae (adapté de Cherwa & Fane, 2011). *Structures des virions du phage PhiX174 (a) et du phage Spv4 (b). Cartes génomiques linéaires des représentants des quatre principaux types de Microviridae (c) et représentation circulaire de la carte génétique de PhiX174.*

parasitaires (*Bdellovibrio bacteriovorus*, *Chlamydiae*), au sujet duquel peu d'informations sont disponibles (Brentlinger *et al.*, 2002). Ainsi, seulement 15 génomes complets de *Microviridae* sont disponibles actuellement au sein du NCBI, alors même que ce type de virus est régulièrement détecté lors d'analyses métagénomiques, que ce soit au sein d'environnements aquatiques (Angly *et al.*, 2006, 2009; Roux *et al.*, 2012) ou du tractus digestif humain (Kim *et al.*, 2011). Nous avons donc réalisé l'assemblage d'un ensemble de viromes publiés, afin de détecter au sein des séquences assemblées de potentiels génomes complets de *Microviridae.* Ces nouveaux génomes pourront ensuite apporter un nouvel éclairage sur cette famille virale, en terme de diversité, organisation du génome, cycle de vie et distribution dans différents écosystèmes.

Article VI

Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads.

Simon Roux, [1,2] Mart Krupovic, [3] Axel Poulet, [1,2] Didier Debroas, [1,2] and François Enault [1,2]

[1]Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", Clermont-Ferrand, France
[2]CNRS, UMR 6023, Laboratoire "Microorganismes : Génome et Environnement", Aubière, France
[3]Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extremophiles, Paris, France

Publié le 11 juillet 2012 dans **Plos One** : 7 (7) e40418

Matériel supplémentaire : Annexe A.7

OPEN ACCESS Freely available online

PLoS one

Evolution and Diversity of the *Microviridae* Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads

Simon Roux[1,2], Mart Krupovic[3], Axel Poulet[1,2], Didier Debroas[1,2], François Enault[1,2]*

1 Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", Clermont-Ferrand, France, **2** Laboratoire "Microorganismes : Génome et Environnement", Aubière, France, **3** Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extremophiles, Paris, France

Abstract

Recent studies suggest that members of the *Microviridae* (a family of ssDNA bacteriophages) might play an important role in a broad spectrum of environments, as they were found in great number among the viral fraction from seawater and human gut samples. 24 completely sequenced *Microviridae* have been described so far, divided into three distinct groups named *Microvirus*, *Gokushovirinae* and *Alpavirinae*, this last group being only composed of prophages. In this study, we present the analysis of 81 new complete *Microviridae* genomes, assembled from viral metagenomes originating from various ecosystems. The phylogenetic analysis of the core genes highlights the existence of four groups, confirming the three sub-families described so far and exhibiting a new group, named *Pichovirinae*. The genomic organizations of these viruses are strikingly coherent with their phylogeny, the *Pichovirinae* being the only group of this family with a different organization of the three core genes. Analysis of the structure of the major capsid protein reveals the presence of mushroom-like insertions conserved within all the groups except for the microviruses. In addition, a peptidase gene was found in 10 *Microviridae* and its analysis indicates a horizontal gene transfer that occurred several times between these viruses and their bacterial hosts. This is the first report of such gene transfer in *Microviridae*. Finally, searches against viral metagenomes revealed the presence of highly similar sequences in a variety of biomes indicating that *Microviridae* probably have both an important role in these ecosystems and an ancient origin.

Citation: Roux S, Krupovic M, Poulet A, Debroas D, Enault F (2012) Evolution and Diversity of the *Microviridae* Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. PLoS ONE 7(7): e40418. doi:10.1371/journal.pone.0040418

Editor: Bas E. Dutilh, Radboud University Nijmegen Medical Centre, NCMLS, The Netherlands

Received April 11, 2012; **Accepted** June 7, 2012; **Published** July 11, 2012

Funding: SR was supported by a PhD grant from the french defense procurement agency (DGA, Direction Générale de l'Armement). MK was supported by the European Molecular Biology Organization (long term fellowship ALTF 347-2010). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: francois.enault@univ-bpclermont.fr

Introduction

Viruses, particularly bacteriophages (viruses of bacteria), are the most numerous biological entities on Earth, retrieved from all sorts of biomes (human body, aquatic ecosystems, soil samples, etc.). Among these, phages with double-stranded DNA (dsDNA) genomes have been the most thoroughly studied [1]. A great deal of data is now available on their diversity [2], relationship with the hosts [3] and their evolution [4]. Such information is still lacking for single-stranded DNA (ssDNA) viruses, which were thought to be secondary actors in environmental viral communities. Yet, it has been recently discovered that these viruses are important members of the virosphere. Indeed, taking into account their modest genome sizes when compared to those of phages with dsDNA genomes, ssDNA viruses were identified in metagenomic datasets from a great variety of ecosystems [5–7]. Their ubiquity led virologists to focus specifically on these viruses [8,9]. Among the ssDNA viruses, the *Microviridae* family is one of the most commonly retrieved.

Microviridae are small icosahedral viruses with circular single-stranded DNA genomes. This family has been thoroughly studied from numerous perspectives – from virion structure and assembly [10–13], to the mechanisms driving their evolution [14], and their stability in environmental conditions [15]. Based on structural and genomic differences, members of this family are further divided into two subgroups : microviruses (genus *Microvirus*) and gokushoviruses (subfamily *Gokushovirinae*) [16]. Very recently, a new tentative sub-family, the *Alpavirinae*, was found through bacterial genome analysis [17], and confirmed by metagenomic analysis [18]. The seven members of the genus *Microvirus* exclusively infect *Enterobacteria* and have been extensively studied through the archetype of this familly, the bacteriophage ΦX174 [10,19]. *Gokushoviruses* are currently known to infect only obligate intracellular parasites, members of bacterial genera *Chlamydia*, *Bdellovibrio* and *Spiroplasma* [20]. Eleven completely sequenced *Gokushovirinae* genomes are currently available: 6 *Chlamydia* phages and 1 genome assembled from seawater viromes [9] are closely-related, whereas *Bdellovibrio* phage phiMH2K [20], *Spiroplasma* phage 4 [21,22], Microvirus ΦCA82 [23], and another genome assembled from a seawater virome [9] are considerably more divergent. Description of *Alpavirinae* is restricted to prophages residing in the genomes of bacteria belonging to two genera of the phylum *Bacteroidetes*: *Prevotella* and *Bacteroides*. The latter study was the first to implicate the *Microviridae* in lysogenization of their hosts and also to associate this virus group with *Bacteroidetes* [17].

Microviridae-like sequences were found in large numbers in different ecosystems, ranging from microbialites [24] to a variety of aquatic environments, with their presence in the GOS dataset [17] and in viral metagenomes [5,6]. Viromes from human stool [18,25] and coral [7] samples were also found to contain *Microviridae*-like sequences. As the known members of the *Microviridae* family exhibit small genomes (3–7 kb), two complete *Microviridae* genomes could be assembled from the Sargasso Sea virome [9].

To gain insights into the diversity of the *Microviridae* viruses in the environment, we reanalyzed a set of previously published viromes by assembling the reads from each of these viromes and then searching the resultant contigs for the presence of complete genome sequences related to *Microviridae*. We were able to assemble 81 complete circular genomes related to members of the *Microviridae* from 95 public viromes. Phylogenetic and genomic organization analyses of these new viruses revealed a new *Microviridae* subgroup (the *Pichovirinae*), enriched the genome collection of *Gokushovirinae*, and, for the first time, confirmed the existence of extrachromosomal complete genomes from *Alpavirinae* virion particles. *Microviridae* core genes could be more thoroughly studied, especially the structure of the major capsid protein. Horizontal gene acquisition events are also documented for the first time in this viral family. Finally, as the viromes analyzed in the current study cover a wide range of ecosystems, the distribution of the new genomes inside the *Microviridae* tree provides a better understanding of both the diversity and the evolution of *Microviridae* family.

Results

Assembly of Complete Genomes

Even though sequences from *Microviridae* are found in a large number of viromes, assembly of complete *Microviridae* genomes was described in only one dataset, the Sargasso Sea virome [26]. Indeed, a consensus sequence of a Chp1-like *Microviridae* was first created by Angly *et al.*, [26] and two complete genomes affiliated to *Microviridae* were assembled in a recent analysis of the same dataset using up-to-date assembly software [9]. To further decipher the evolution and distribution of this family, we assembled all available public viromes (with a threshold of 98% identity on 35 bp) and screened them for the presence of complete *Microviridae*-like circular genomes. As a result, 81 contigs representing putative complete *Microviridae* genomes were obtained from 25 out of the 95 viromes tested (Table S1). Out of these 81 new genomes, 15 new *Microviridae* were assembled from freshwater virome reads, 2 from a marine sample, 2 from microbialites, 1 from coral, 2 from human lung and 59 from human gut samples. Obviously, the number of *Microviridae* genomes assembled from a virome depends on the length of the sequences (Table S1) but the assembly also depends on the relative abundance of a given virotype in the sampled ecosystem. The sizes of the different genomes generated were quite homogeneous (Table S2), with the smallest genome being 3,989 bp-long and the longest 6,723 bp. This size range is consistent with the genome sizes of known *Microviridae* (between 4.4 and 6.1 Kb).

In order to detect potential new prophages (viral genomes integrated into bacterial chromosomes), the newly assembled *Microviridae* sequences were used as queries in searches against the complete bacterial genomes from the NCBI and the Human Microbiome Project [27] databases. The five previously described complete prophages [17] were fully retrieved, alongside a new one, detected in the recently sequenced *Prevotella multiformis* strain, highly similar to other *Microviridae* prophages detected in *Prevotella*.

Phylogeny and Genome Organization of the Microviridae Family

In order to gain insight into the diversity of the *Microviridae* family and its genome evolution, a phylogenetic tree was computed from the major capsid protein (VP1) sequences and correlated with the corresponding genome maps (Fig. 1).

Phylogenetic analysis using the VP1 protein. Four well-supported clades (bootstrap values higher than 75) are formed on the VP1 phylogenetic tree: three of these correspond to the previously described taxonomic groups (*i.e.* genus *Microvirus*, and subfamilies *Gokushovirinae* and *Alpavirinae*), while the fourth one is exclusively composed of genomes generated in this study. The assembled viromes considerably expanded the available complete viral genome pool for *Gokushovirinae* and *Alpavirinae*, while not a single new virus was affiliated to the genus *Microvirus*. These four groups are further described below:

- The Enterobacteria phage group (the yellow group in Fig. 1) is exclusively composed of known members of the genus *Microvirus*. These phages are clearly separated from the rest of the *Microviridae*, consistent with the current ICTV taxonomy, where they form a distinct genus within the *Microviridae* family.
- The **Alpavirinae group** (the pink group in Fig. 1) includes the recently described *Alpavirinae* and 33 newly assembled *Microviridae* genomes. This group can be divided into three subgroups : two subgroups are exclusively composed of new genomes from the human gut flora. The first subgroup consists of 18 related genomes, while the second one encompasses three more divergent viruses Human_feces_B_007, Human_-gut_21_005, and Human_feces_C_010. The third subgroup is composed of the 6 prophages, a new virus from a human lung sample and 11 viruses from the human gut samples. Notably, prophages from *Bacteroides* are separated from those inserted into the genomes of *Prevotella*. Consistently, the new *Prevotella* prophage (*Prevotella multiformis*) detected in this study is more similar to prophages from other *Prevotella* genomes. Interestingly, a new *Microviridae*, generated from free viral particles from a human lung sample, is interspersed on the tree among these *Prevotella* prophages. *Bacteroides* prophages form a monophyletic group with 11 new genomes sampled from human gut flora.
- The ***Gokushovirinae* group** (the blue group in Fig. 1) consists of all known *Gokushovirinae* and 42 newly assembled *Microviridae* genomes. These new viruses come from different ecosystems : 27 from the human gut flora (9 different individuals), 12 from freshwater lakes (11 from Lake Bourget and 1 from Lake Pavin), 1 from marine environment (JCVI_001, sampled from Chesapeake Bay), 1 from human lung (SectLung2LLL_002) and 1 from microbialites (68_Microbialites_003). This group can be internally divided : *Gokushovirinae* assembled from aquatic samples are most closely related to *Bdellovibrio* phage ΦMH2K, whereas sequences assembled from human gut samples are divided into two subgroups, one around *Spiroplasma* phage 4 and Microvirus ΦCA82, and another group close to *Chlamydia* phages.
- The **new group** (the green group in Fig. 1) is composed exclusively of new viruses assembled from metagenomic sequences. This group contains 3 genomes from two different freshwater lakes, 1 from marine water, 1 virus associated with coral microbiota and 1 from microbialites. As this group is separated from the already defined groups, we propose to name it ***Pichovirinae*** (*Picho* : small in Occitan).

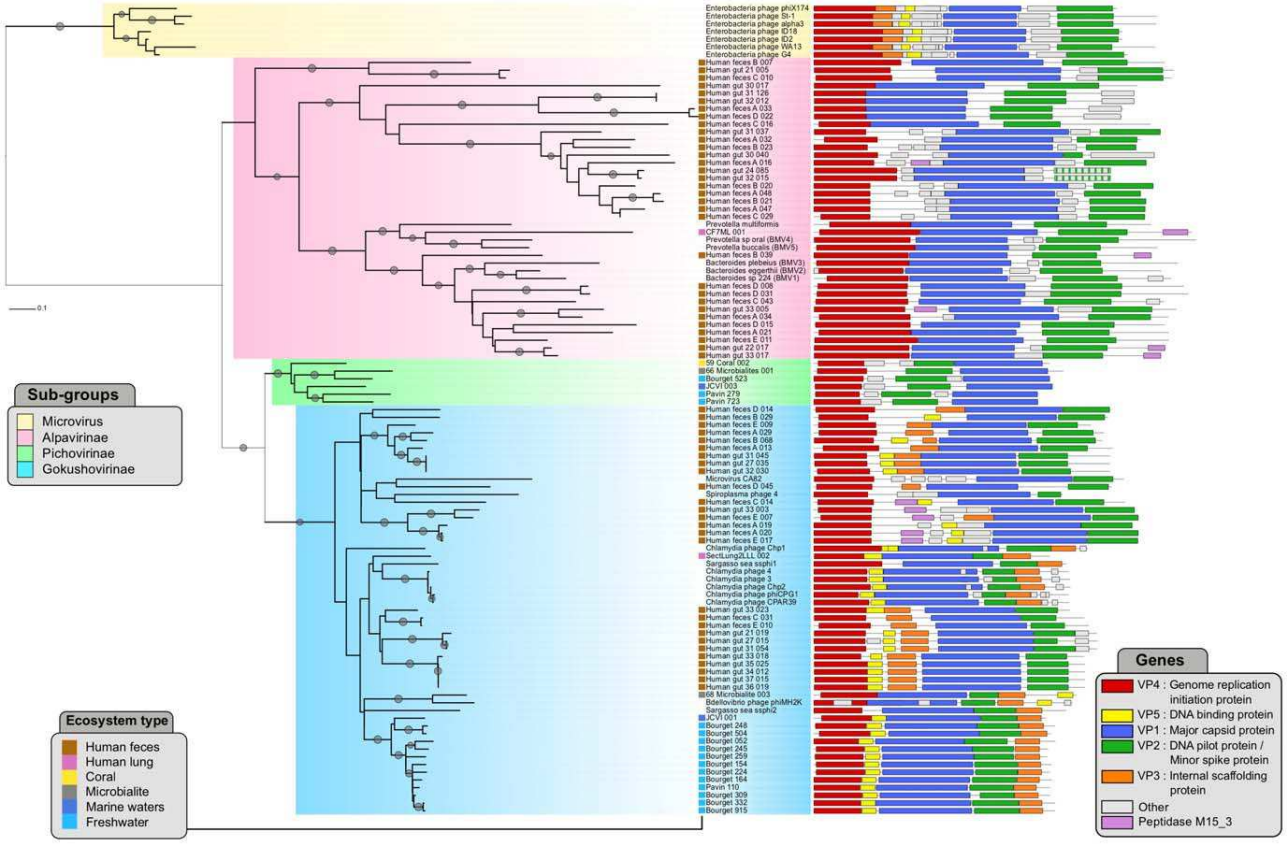


Figure 1. Phylogenetic tree drawn from the major capsid protein multiple alignment. Linearized genomes are represented for each virus. The open reading frames in each genome are color-coded following the nomenclature used for *Chlamydia* phage genomes (i.e.VP1 : major capsid protein, VP2 : DNA pilot protein, VP3 : internal scaffolding protein, VP4 : genome replication initiation protein, and VP5 : DNA binding protein). Striped-colored genes encode proteins possessing features characteristic of VP2 proteins, but displaying no significant sequence similarity, as assessed by BLAST. The four *Microviridae* subgroups are highlighted on the tree. Bootstrap scores greater than 80 are marked with gray dots.
doi:10.1371/journal.pone.0040418.g001

Genome analysis of Microviridae. All 81 new *Microviridae* genomes were composed of 3 to 9 predicted genes. These gene numbers are consistent with the known reference genomes, and were similar for each subgroup. The *Microviridae* core genes (encoding the major capsid protein VP1, minor spike or pilot protein VP2 and replication initiation protein VP4) are detected in all *Microviridae* genomes but two (Fig. 1). The average genome size of the four different sub-families was found to be significantly different (Fig. S1, one-way ANOVA, p-value 2.2e-16). Notably, *Microvirus* and *Alpavirinae* genomes are longer than those of *Pichovirinae* and *Gokushovirinae*. Furthermore, the genomic organization of each of the 4 sub-families is specific. Indeed, the genome organizations are conserved within the 4 sub-families but different between the sub-families (Fig. 1).

Genomes of *Alpavirinae* display a reduced content of *Microviridae* conserved genes, with only three genes (for proteins VP1, VP2 and VP4) being significantly similar to those of *Microviridae* from other genera/subfamilies. Two new genomes (Human_gut_24_085 and Human_gut_32_015), sampled from two different individuals, lack an ORF significantly similar to VP2, but instead possess a similarly-sized ORF at a position equivalent to that occupied by VP2 in all other members of the *Alpavirinae* (Striped-colored genes on Fig. 1). These ORFs are likely to be highly divergent VP2-coding genes. Consistently, the putative products of both ORFs display features characteristic of all VP2-like proteins. Namely, both gene products possess predicted N-terminal transmembrane domains and coiled-coil regions. Although no other microviral genes could be detected within the *Alpavirinae* prophages and related assembled genomes using sequence-based searches, it has been previously suggested that VP3-like scaffolding proteins might be encoded transcriptionally downstream from the VP1-encoding genes [17]. Indeed, most of the *Microviridae* from the *Alpavirinae* group possess unassigned ORFs that might encode an equivalent of gokushoviral protein VP3.

Genomes of *Gokushovirinae* share the same gene content (presence of ORFs significantly similar to VP1, VP2, VP3, VP4 and VP5), with the exception of *Spiroplasma* phage 4, Microvirus ΦCA82, Sargasso sea phage ssΦ2 and 12 new genomes from human gut, for which VP3 and/or VP5-like genes could not be identified using standard sequence analysis methods. However, upon a closer examination using a sensitive profile-profile comparison algorithm implemented in FFAS03 [28], an ORF potentially encoding a homologue of VP3 has been identified in 4 of these genomes from human gut (Human_Gut_33_003, Human_Feces_A_019, Human_Feces_A_020, Human_Feces_E_017; hit to *Chlamydia* phage Chp2 scaffold protein VP3 superfamily, pfam id : PF09675; FFAS03 mean score: −29.2). An internal separation of the *Gokushovirinae* assembled in this study into two subgroups can be deduced from the gene order conservation within these subgroups, consistently with the phylogenetic

information. A first subgroup is found near *Bdellovibrio* phage ΦMH2K, and encompass only genomes assembled from aquatic environments (Lake Bourget, Lake Pavin, and JCVI_001, sampled from Chesapeake bay). This subgroup displays a specific gene order : VP4-VP5-VP1-VP2-VP3 (Fig. 1). Genomes assembled from human gut present a different gene order (VP4-VP5-VP3-VP1-VP2), and do not form a monophyletic group within the *Gokushovirinae*. Yet, the low bootstraps scores point towards the possibility that the internal branching within this group might change once more gokushoviral sequences become available. Finally, two exceptions have to be noted : first, a sequence assembled from Human Lung sample (SectLung2LLL_002) is found near the known *Chlamydia* phages, and presents a gene order similar to the aquatic *Gokushovirinae* assembled in this study (Fig. 1, VP4-VP5-VP1-VP2-VP3); second a genome assembled from Microbialites (68_Microbialites_003) displays the same gene order as *Bdellovibrio* phage ΦMH2K, and is related to this phage in the tree with a significant bootstrap support.

The gene composition of *Pichovirinae* genomes is similar to that of *Alpavirinae*: significant sequence similarity with known references are only detected for the three major genes (for VP1, VP2 and VP4). Nevertheless, *Pichovirinae* genomes are the only ones within the *Microviridae* family where the gene order of the core genes is altered: whereas all *Microviridae* present a VP4 - VP1 - VP2 organization, the gene order in all *Pichovirinae* is VP4 - VP2 - VP1 (Fig. 1).

Detailed Analysis of the Conserved Microviridae Proteins

Major capsid protein (VP1). Virions of microviruses and gokushoviruses display distinct structural features and molecular composition; although both possess icosahedral capsids composed of 60 copies of the major capsid protein, MCP (F and VP1, respectively), only those of microviruses are decorated with pentameric major spike protein complexes positioned at each of the 12 five-fold vertices [11]. Electron cryo-microscopy (cryo-EM) study of the SpV4 virions revealed that gokushoviruses instead possess 55 Å-long 'mushroom-like' protrusions located at the 3-fold symmetry axes of their capsids [21]. These protrusions are formed by insertion loops coming from three subunits of the VP1 protein (Fig. 2A), and were suggested to participate in receptor recognition and binding on the host cell surface. The protrusion-forming insertion is not present in the MCPs of ΦX174-like microviruses and is largely accountable for the size differences between the MCPs of microviruses and gokushoviruses (Fig. 2B).

Comparison of the MCP sequences from the four subgroups of the *Microviridae* revealed that the average identity at the protein level varies from 40 to 80% within a group, and from 20 to 40% between the groups, with an exception of the *Enterobacteria*-infecting phages, which generally present no significant sequence identity with MCPs from other clades of the *Microviridae* (Fig. S2). Analysis of the VP1 size variation including the newly discovered members of the *Microviridae* revealed that the average MCP size in the four different sub-families is significantly different (one-way ANOVA, p-value 1e-43). ΦX174-like microviruses possess smaller MCPs (average length 427 aa), while those of gokusho-, alpa- and pichoviruses are significantly larger (Fig. 2B). Notably, among the latter three groups, pichoviruses possess the smallest MCPs (average length 512 aa), while the MCPs of alpaviruses are the largest (average length 630 aa) and also the most variable in terms of size (ranging from 541 to 780 aa). Multiple sequence alignment of VP1 homologues revealed that the MCP size difference is a result of variation in the number and size of insertions in VP1-like proteins (Fig. S3).

To gain insights into possible architecture of viruses from the newly identified groups *Alpavirinae* and *Pichovirinae* and to understand the effect that the insertions within their MCPs might have on virion architecture, we constructed three-dimensional VP1 models for representative viruses. These were compared to the available X-ray structure of ΦX174 protein F (PDB ID:2BPA; [11]) and the cryo-EM-based model of SpV4 protein VP1 (PDB ID:1KVP; [21]). Structural modeling and model quality assessment are described in Materials and Methods. Comparison of the structural models (Fig. 2C) revealed that VP1 homologues from viruses belonging to all four groups of the *Microviridae* possess a conserved eight-stranded β-barrel core (also known as viral jelly-roll; [29]) and all, but ΦX174-like microviruses, possess an extended loop that forms a mushroom-like protrusion in SpV4. Consequently, it is likely that virions of gokusho-, alpa- and pichoviruses, unlike those of microviruses, possess characteristic receptor-binding spikes at the three-fold axes of the icosahedral capsids (Fig. 2A).

Further analysis has revealed that the size of the putative receptor-binding spike-forming insertions differs between different subgroups of *Microviridae* (Fig. S4A): the shortest are found in pichoviruses (average length 60 aa), while those of alpaviruses are the longest (average length 110 aa). The insertion length also varies considerably within *Gokushovirinae* (from 53 to 114 aa) and *Alpavirinae* (from 45 to 209 aa). Interestingly, this variation appears to be ecosystem- rather than virus subgroup-dependent. The insertion length variation was much less pronounced for VP1 proteins of viruses residing in aquatic environments (including both gokushoviruses and pichoviruses) than it was for viruses from human samples (gokushoviruses and alpaviruses) (Fig. S4B). The same tendency was also true for the full length MCPs, with VP1s from human samples being larger (average length 604 aa) than those of viruses thriving in aquatic environments (average length 533 aa). It therefore appears that evolution towards acquisition of insertions within the MCPs of viruses isolated from human samples might be driven by the need to cope with additional factors (e.g., immune system, low pH of the human gastrointestinal tract, etc.) that are not present in aquatic environments.

Distinct members of the *Alpavirinae* (a group exclusively associated with human samples; Fig. 1) possess insertions at different locations within their VP1 proteins, suggesting that VP1 proteins within this *Microviridae* subgroup are indeed evolving rapidly. Such species-specific insertions were found to be up to 231 aa-long (Human_feces_B_007). In order to verify whether such extensive insertions would not interfere with normal virion formation, we fitted our three-dimensional model of the alpaviral VP1 into the pseudoatomic model of SpV4 (PDB ID:1KVP) and mapped the location of all the insertions exceeding 15 aa. We identified 6 MCP hot-spots where large insertions were tolerated in alpaviruses (Fig. S5). Notably, all of these insertions occurred in the loop regions of the MCP facing outwards from the virion surface and are therefore expected not to affect virion assembly. Interestingly, the 231 aa-long insertion in the MCP of Human_-feces_B_007 is predicted to be rich in β-strands and is likely to fold into an independent domain. Peculiarly, the major spike protein G of ΦX174-like microviruses, which forms protrusions at the virion five-folds of these viruses is also rich in β-strands. Unfortunately, the sequence of this insertion in the MCP of Human_feces_B_007 does not share significant similarity with proteins in extant databases and its provenance therefore remains obscure.

Replication protein (VP4). The replication protein is highly variable in length, as some microphages possess long replication genes (namely *Alpavirinae* assembled from prophages and the associated virions, but also *Chlamydia* phage Chp1, and Sargasso

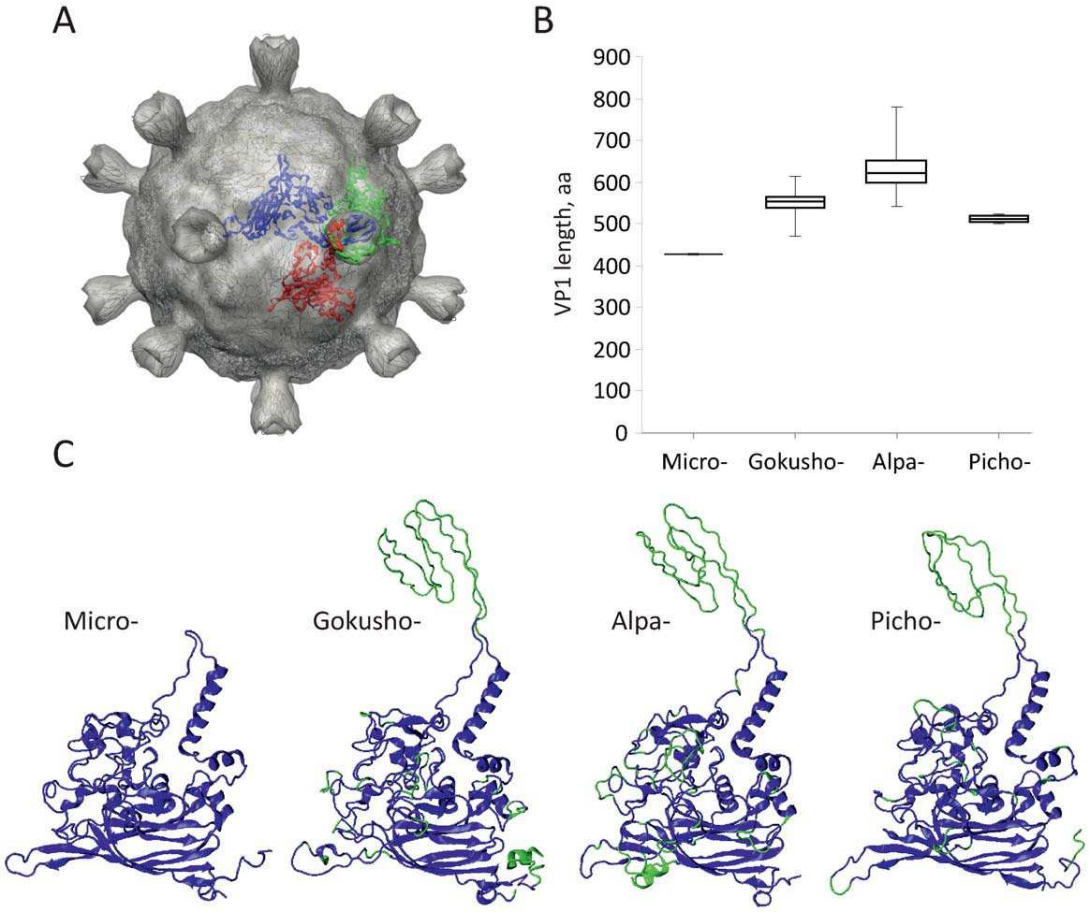


Figure 2. Major capsid protein (MCP) variation within the *Microviridae*. (A) Three-dimensional model of the SpV4 virion (PDB ID:1KVP). Three capsomers donating long insertion loops to form the 'mushroom-like' protrusions at the three-fold axes of symmetry of the icosahedral capsid are highlighted in blue, green, and red. (B) A boxplot illustrating the variation of MCP sizes between the four subgroups of the *Microviridae*. (C) Three-dimensional models of the MCPs from viruses representing the four subgroups of the *Microviridae*: *Microvirus* (ΦX174 protein F; PDB ID:1CD3), *Gokushovirinae* (SpV4 VP1, PDB ID:1KVP), *Alpavirinae* (*Prevotella bucalis* prophage BMV5 protein VP1; GI:282877220), *Pichovirinae* (Pavin_279 protein VP1). The insertions within the VP1 proteins of gokusho-, alpa- and pichoviruses relative to the F protein of ΦX174 are highlighted in green.
doi:10.1371/journal.pone.0040418.g002

sea phage ssphi1). Nevertheless, the three conserved motifs of superfamily I rolling cycle replication proteins are all conserved (Fig. S6), suggesting that these proteins are likely to be functional. High levels of sequence identity are detected within all *Enterobacteria* phages sequences, as well as within *Pichovirinae and Gokushovirinae* (Fig. S7). Conversely, replication proteins from *Alpavirinae* are considerably less conserved within the group. Globally, the similarity between VP4 sequences for any given pair of viruses is lower than the one for the VP1 sequences from the same pair of viruses.

DNA pilot protein (VP2). The last gene retrieved in all *Microviridae* genomes to date codes for the pilot protein (VP2 in *Gokushovirinae*, Minor spike protein H in *Enterobacteria* phages). The ΦX174 protein H is a multifunctional structural protein (12 copies per virion) required for piloting the viral DNA into the host cell interior during the entry process, and *de novo* synthesis of protein H is required for efficient production of other viral proteins [30–32]. However, the full functional potential of this protein remains to be elucidated. At the first glance, VP2 appears to be more divergent than the MCP or replication protein: significant sequence similarity is only detected within sequences of the same subgroup (*Gokushovirinae*, *Enterobacteria phages*, *Alpavirinae* and *Pichovirinae*, Fig. S8). Strikingly however, similarity between VP2 proteins from more closely related viruses often equals or even exceed the similarity observed between their major capsid or replication proteins. This is, for example, the case for *Chlamydia* phages (with the exception of the highly divergent *Chlamydia* phage 1; Fig. S8), and *Enterobacteria*-infecting *phages*. This perplexing host-dependent pattern of VP2 conservation raises the possibility that the evolution and function(s) of this protein might be tightly linked to the identity of the host.

Horizontal Acquisition of New Genes

It has been previously suggested that genes encoding novel functions in microviral genomes emerge from pre-existing genomic regions through accumulation of point mutations [19,33]. This conclusion has been supported by the lack of identifiable cases of horizontal acquisition of new genes by the *Microviridae*. Analysis of the complete microviral genomes assembled in this study has unexpectedly revealed 11 genes from human-associated *Microviridae* (5 *Gokushovirinae* and 6 *Alpavirinae* - 10 from human gut and 1 from human lung; Fig. 1) encoding a putative peptidase of the M15_3 family (Pfam Id : PF08291). M15 family peptidases are widespread in bacteria and are involved in

cell wall biosynthesis and metabolism; they catalyze hydrolytic cleavage of the amide bond within peptide bridges that cross-link glycan strands of the bacterial cell wall [34].

The closest homologues detected by BLAST for these 11 genes are from bacterial genomes, except for Human_feces_A_016, for which the closest homologue is found in a tailed dsDNA phage genome (Table S3). Half of the *Alpavirinae* peptidase genes are affiliated to *Bacteroidetes* and a *Bacteroidetes* phage, the three others are associated with *Burkhorderiales* (*Leptothrix* and *Collimonas*). *Gokushovirinae* peptidases are affiliated to *Firmicutes* : *Faecalibacterium prausnitzii* (4 of 5), and *Gamma-proteobacteria* (*Providencia alcalifaciens*). These closest homologues of the microviral proteins are found next to phage-like genes in several bacterial genomes (Fig. 3A). For example, phage-like integrase genes are proximal to the M15 peptidase genes in *Bacteroides vulgatus ATCC* 8482 and *Bacteroides vulgatus PC510* genomes, indicating a likely phage origin for these genomic regions. Consistently, the peptidase gene from *Providencia alcalifaciens DSM30123* genome is retrieved within a complete prophage region, and next to a putative holin gene. The peptidase gene from *Faecalibacterium prausnitzii* M21/2 is present within a three-gene cassette, with all three genes having homologues in bacteriophages (Fig. 3A). Notably, besides the M15 peptidase, the cassette includes a putative holin gene (hit to *Lactococcus* phage ul36.t1, ABD63797; 47% identity, E = 1e-20) and a gene of unknown function, with a homologue present as part of the lysis gene cluster in *Streptococcus* phage 858 (Fig. 3A). All this indicates that peptidase genes are likely to be frequently exchanged between viruses and their hosts, probably through prophage integration. Interestingly, 5 of the 11 peptidase genes detected in *Microviridae* genomes are adjacent to an "unknown" predicted ORF, which displays no sequence similarity to proteins in the extant databases. Peculiarly, this unknown gene (153 codons) in the Human_gut_33_003 genome is encoded on the complementary strand. Such orientation is highly unusual; to our knowledge, no other cases of complementary strand genes have been reported in *Microviridae*.

In order to shed light on the evolutionary event(s) leading to the acquisition of M15 peptidase genes by *Microviridae* phages, a phylogenetic tree was computed from a multiple alignment including the *Microviridae* peptidases and their closest homologues in both bacterial and viral genomes (Fig. 3B, Fig. S9). The topology of the peptidase tree is consistent with the VP1-derived tree of *Microviridae*, with a clear separation between the peptidase genes from the *Alpavirinae* (highlighted in pink) and the five peptidase genes from *Gokushovirinae* (in blue). Within the *Alpavirinae*, the peptidase tree topology can be associated with the location of the peptidase gene integration within these viral genomes. Peptidases genes are inserted in two different positions in *Alpavirinae* genomes : between VP2 and VP4 (CF7ML001, Human_feces_B_039, Human_gut_22_017 and Human_gut_33_017), and between VP4 and VP1 (Human_feces_A_016 and Human_gut_33_005) (Fig. 1). Consistently, these two groups are retrieved on the peptidase tree: Human_feces_A_016 and Human_gut_33_005 are found near the *Fusobacterium* gene within the *Bacteroidetes* group, whereas the other *Alpavirinae* peptidases are retrieved at the base of the *Bacteroidetes* group. Within the *Gokushovirinae*, the Human_feces_C_014 peptidase is separated from the rest of gokushoviral peptidases, similarly to the topology of the VP1 tree (Fig. 3B, Fig. 1). In the peptidase tree, the closest neighbors of the *Gokushovirinae* peptidases are from *Firmicutes* (*Faecalibacterium)* and *Proteobacteria* (*Ahrensia*, *Providencia*). The most likely explanation for such clustering is that peptidase genes were horizontally acquired by several members of the *Microviridae* on multiple occasions. The presence of dsDNA phage peptidases near the *Microviridae* sequences on the tree suggests that ds and ssDNA phages might be engaged in gene exchange, either directly during a co-infection or via infection of a prophage-bearing host cell.

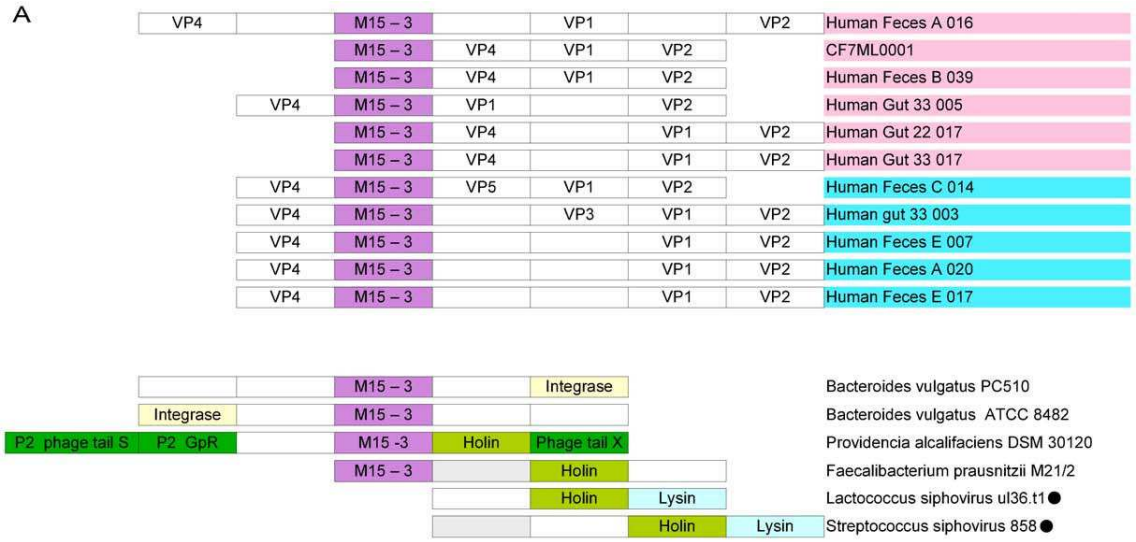
Finally, the possibility of horizontal gene transfer between *Microviridae* genomes was investigated in light of this acquisition of a peptidase gene by several human gut *Microviridae*. For that, we performed a phylogenetic analysis of the two most conserved *Microviridae* proteins, VP1 and VP4, for each of the four subgroups (*Microvirus, Alpavirinae, Gokushovirinae* and *Pichovirinae*). No signs of recent gene transfer event could be detected, confirming the hypothesis that gene transfer between *Microviridae* are rare, even within the temperate members of the group [14,17,19].

Microphages Diversity in Environment

Biogeographic pattern and Microviridae dispersal. The wide distribution of samples from which complete *Microviridae* genomes could be assembled (Table S1) makes it possible to analyze the repartition of the different *Microviridae* subgroups among the different geographic sites sampled, and the different types of environments studied. Remarkably, *Microviridae* genomes could be assembled from all but hypersaline and hyperthermophilic types of samples. Very similar genomes are retrieved from geographically remote sampling sites, both in aquatic medium (for examples JCVI_001 from North America is closely related to Bourget_248 and Bourget_504, from France) and in human microbiome : genomes noted as "Human feces", sampled in South Korea [18], are not very different from the "Human gut" genomes, sampled in North-America [25]. This wide distribution and absence of biogeographic pattern is likely to reflect an ancient origin for *Microviridae*, which would have colonized a wide range of habitats, from human microbiome to seawater, freshwater, and sedimentary structures like Microbialites.

Nevertheless, assembling complete *Microviridae* genomes from random environmental sequences requires a large number of reads, especially for viromes composed of reads not exceeding 100 bp. Thus, only 3 complete *Microviridae* genomes have been generated from the three viromes with reads of ~100 bp (from a total of 41 such viromes in the dataset, Table S1). To gain further insights into the diversity and patterns of distribution of *Microviridae*, a database of major capsid protein (VP1) sequences was built encompassing all published and newly assembled *Microviridae*. These sequences were used to search for VP1 homologues in the unassembled virome reads.

Viral metagenome sequences similar to VP1. A total of 498 sequences were found to be significantly similar (BLASTx bit score greater than 50) to the VP1 protein of at least one of the *Microviridae* complete genomes. These 498 metagenomic sequences span 36 of the 95 viromes, from 5 different types of ecosystem (human microbiome, other eukaryote, seawater, freshwater, microbialites). *Microviridae* remains undetected in hypersaline and hyperthermophilic environments (Fig. 4). In order to analyze the dispersal of each *Microviridae* subgroup, the presence of each subgroup was checked in the 36 viromes containing *Microviridae* sequences. The *Gokushovirinae* subgroup is the most widespread among *Microviridae* (28 viromes out of 36), and is found in all *Microviridae*-containing biomes (Fig. 4). *Pichovirinae* are less frequently detected (12 viromes), but are also retrieved from different types of biomes. On the contrary, *Alpavirinae* are exclusively detected in human sample viromes (16 samples). Finally, only one sequence affiliated to the genus *Microvirus* was detected in a seawater virome. Yet, the low BLAST bit score (50.1) and the fact that this is the only microvirus-like sequence retrieved indicates that microviruses are likely to be extremely rare in such environments.


A
VP4
M15 – 3
VP1
VP2
Human Feces A 016
M15 – 3
VP4
VP1
VP2
CF7ML0001
M15 – 3
VP4
VP1
VP2
Human Feces B 039
VP4
M15 – 3
VP1
VP2
Human Gut 33 005
M15 – 3
VP4
VP1
VP2
Human Gut 22 017
M15 – 3
VP4
VP1
VP2
Human Gut 33 017
VP4
M15 – 3
VP5
VP1
VP2
Human Feces C 014
VP4
M15 – 3
VP3
VP1
VP2
Human gut 33 003
VP4
M15 – 3
VP1
VP2
Human Feces E 007
VP4
M15 – 3
VP1
VP2
Human Feces A 020
VP4
M15 – 3
VP1
VP2
Human Feces E 017
M15 – 3
Integrase
Bacteroides vulgatus PC510
Integrase
M15 – 3
Bacteroides vulgatus ATCC 8482
P2 phage tail S
P2 GpR
M15 -3
Holin
Phage tail X
Providencia alcalifaciens DSM 30120
M15 – 3
Holin
Faecalibacterium prausnitzii M21/2
Holin
Lysin
Lactococcus siphovirus ul36.t1
Holin
Lysin
Streptococcus siphovirus 858

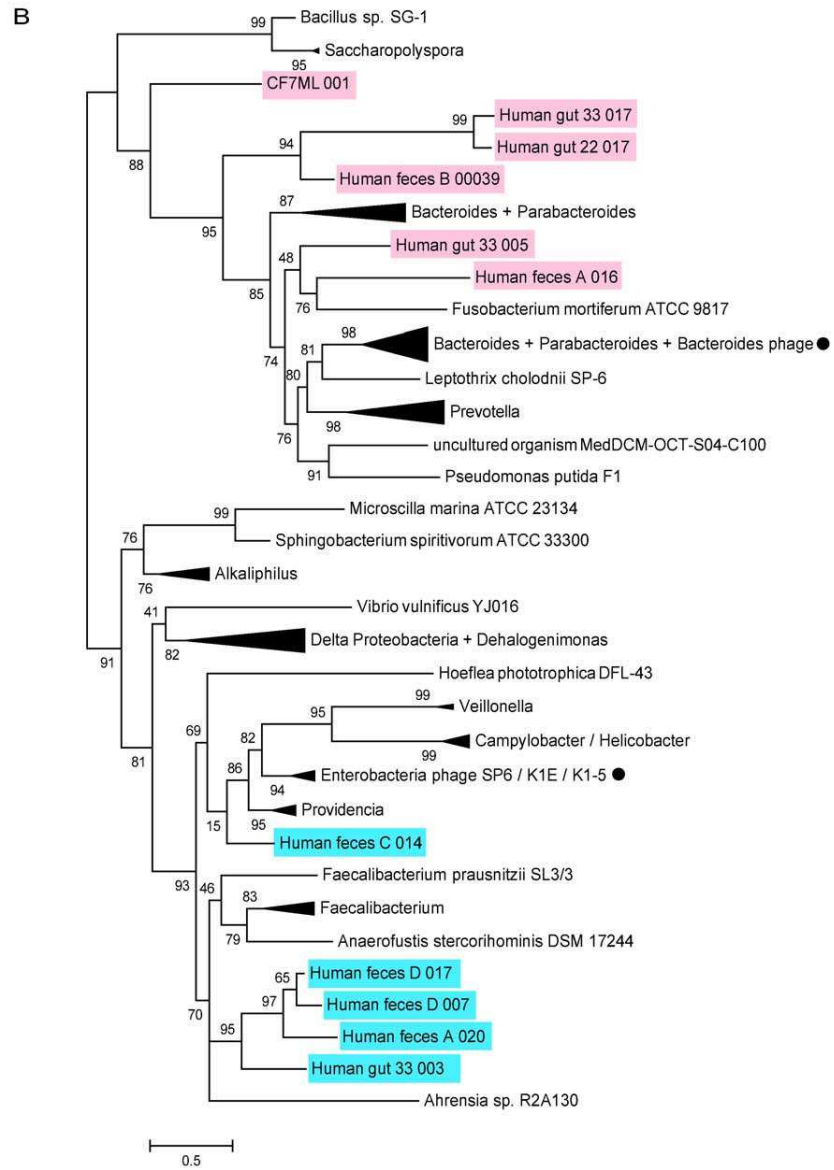

B
Bacillus sp. SG-1
Saccharopolyspora
CF7ML 001
Human gut 33 017
Human gut 22 017
Human feces B 00039
Bacteroides + Parabacteroides
Human gut 33 005
Human feces A 016
Fusobacterium mortiferum ATCC 9817
Bacteroides + Parabacteroides + Bacteroides phage
Leptothrix cholodnii SP-6
Prevotella
uncultured organism MedDCM-OCT-S04-C100
Pseudomonas putida F1
Microscilla marina ATCC 23134
Sphingobacterium spiritivorum ATCC 33300
Alkaliphilus
Vibrio vulnificus YJ016
Delta Proteobacteria + Dehalogenimonas
Hoeflea phototrophica DFL-43
Veillonella
Campylobacter / Helicobacter
Enterobacteria phage SP6 / K1E / K1-5
Providencia
Human feces C 014
Faecalibacterium prausnitzii SL3/3
Faecalibacterium
Anaerofustis stercorihominis DSM 17244
Human feces D 017
Human feces D 007
Human feces A 020
Human gut 33 003
Ahrensia sp. R2A130
0.5

Figure 3. Genomic context and phylogenetic analysis of the Peptidase M15 *Microviridae* sequences. (A) Organization of the Peptidase M15_3 region in the 11 newly assembled *Microviridae* genomes. The regions encompassing homologous peptidase genes in three bacterial and two phages (noted with a black circle) genomes are also shown. *P2 GpR* stands for the P2 phage tail completion protein. (B) Maximum-likelihood tree computed from the multiple alignment of peptidase M15 sequences of the *Microvidae* and their closest homologues in viral and bacterial genomes. Bootstrap support values are indicated on each node. A fully expanded view of this tree is available as Fig. S9.
doi:10.1371/journal.pone.0040418.g003

Discussion

Microphages are progressively retrieved from a broad range of environmental samples from various locations. In this study, the focus on this family through a search in viral metagenomes made it possible to describe a new subgroup of *Microviridae* and considerably expanded the existing knowledge on genome evolution, diversity and environmental distribution of this viral family. Using already published viromes that have never been analyzed for the presence of complete *Microviridae* genomes, this study more than tripled the number of complete genomes available for this family.

Technical Issues and Potential Bias of the de novo Metagenomic Assembly of Viral Genomes

Several points have to be discussed regarding both bioinformatical and biological issues in order to better understand the results obtained in this study. First, sequences reconstructed here from metagenomic data represent consensus sequences of individual microviral populations and thus, DNA sequence variations within each population are masked. Nevertheless, the assembly criteria used here (98% similarity on 35 bp) are considered stringent enough to gather only sequences from the same viral species [26]. Such criteria both limit sequence variations within each assembled genome and mask pyrosequencing errors. Second, even if *Microviridae* seems more abundant and frequently retrieved in particular ecosystem types (as in the human gut), all biomes were not evenly sampled. Indeed, the published viromes used in this study were generated independently and with different sequencing technologies. For instance, viromes from the human gut contain approximately 2 times more base pairs than viromes from sea water, and 6 times more than those from hypersaline ponds. Third, the methodology used to generate a virome greatly influences the type of viruses which will be retrieved. Viromes prepared through LASL (Linker Amplified Shotgun Library) are not supposed to contain any ssDNA genomes as this technique only recover dsDNA fragments. Not surprisingly, no *Microviridae* were detected in the two viromes prepared through LASL in this study, both sampled from hyperthermophilic environments. Last, the quantitative importance of ssDNA viruses in general, and of *Microviridae* more specifically, remains an open question. The abundance of *Microviridae* was first considered to be higher than predicted when the affiliation of virome reads was normalized by the mean genome length, *i.e.* when viruses were compared in terms of "number of viral particles in the original samples" [5]. Nevertheless, this quantitative importance was balanced by a potential bias of the whole-genome amplification methodology, which would preferentially amplify small circular DNA templates [35]. Although the real relative abundance of the *Microviridae* is still unknown, the fact remains that a considerable part of reads in numerous viromes could be affiliated to *Microviridae*, leading to the assembly of 81 complete *Microviridae* genomes.

Microviridae is a Coherent Family of ssDNA Phages

The new genomes assembled in this study confirmed that *Microviridae* is a consistent and homogeneous viral family. All but two genomes contained ORFs significantly similar to the three core genes of *Microviridae* (the major capsid protein VP1, the replication protein VP4, and the DNA pilot protein VP2). Moreover, the division of the *Microviridae* family into sub-families, which has been previously presented [20], was both confirmed and complemented by this study. The microphage diversity deduced from the analysis of the environmental samples confirmed the distinction between *Enterobacteria*-infecting phages (genus *Microvirus*), and the other known *Microviridae*. Members of the genus *Microvirus* are considered as *Microviridae* archetypes, and are the most studied *Microviridae* so far, yet none of the genomes assembled

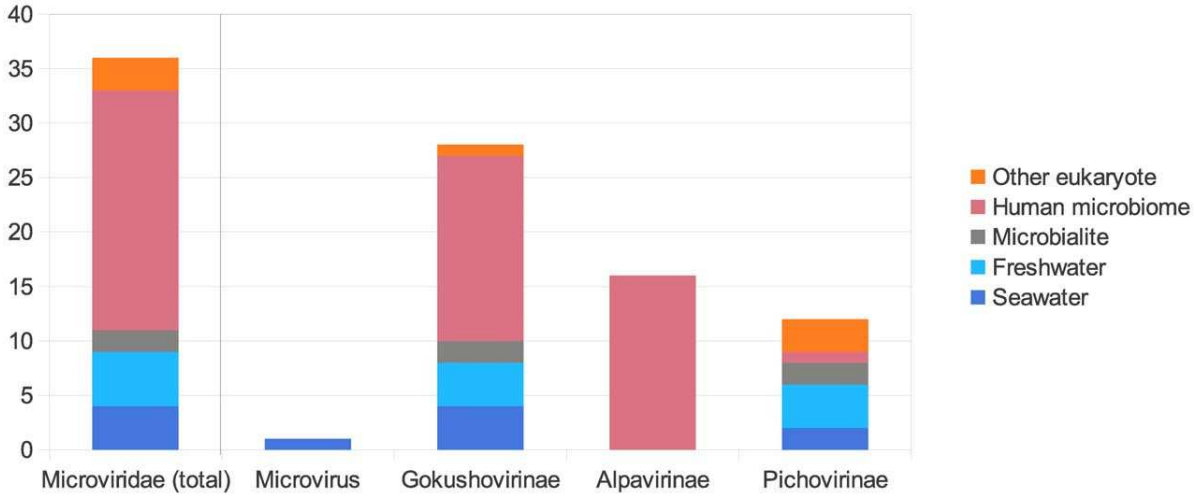


Figure 4. Abundance and distribution of VP1-like sequences in the environment. The number of viromes and the origin of the samples used for virome preparation are indicated. VP1 sequences were affiliated by best BLAST hit against a database including VP1 sequences from both the previously published *Microviridae* genomes and the complete genomes assembled in this study.
doi:10.1371/journal.pone.0040418.g004

in this study is associated with this genus. This apparent paradox might be due to culture bias, as *Microviruses* are cultivated on *E. coli* strains, the most widely used and studied prokaryotic model organism. The present study indicates that this genus is rare among the biosphere, and likely constitute a very specific type of *Microviridae* especially in terms of gene content and capsid structure. The other *Microviridae* appear to be internally divided into three main subgroups, namely *Gokushovirinae*, *Alpavirinae*, and a new subgroup, which we propose to name *Pichovirinae*. This new clade appear to be more related to *Gokushovirinae* than to *Enterobacteria* phages, and could represent an intermediate group, like *Alpavirinae*, which could fill the gap between the two currently recognized sub-families. Moreover, this new group has a unique genome organization, as VP2 is located between VP1 and VP4 for all the genomes in this group. Based on this synteny break, and the fact that *Pichovirinae*-like genomes have been assembled from very different samples (namely seawater, freshwater, microbialites, and coral samples), this clade has probably diverged from the common *Microviridae* ancestor a long time ago. Consistently, significant sequence similarity between pichoviruses and the other *Microviridae* is confined to the major capsid protein VP1 and the replication protein VP4, while the pilot proteins (VP2) displays only limited sequence similarity to corresponding proteins from the *Alpavirinae* prophages.

This taxonomic structure of the *Microviridae* family confirms that the microphage diversity was under-estimated. The detection of sequences homologous to *Pichovirinae* and *Gokushovirinae* in very different environments (Fig. 4) suggests that the viruses giving rise to these clades have probably diverged from their common ancestor in a distant past. The latter possibility is supported by the differential gene order conservation in the *Pichovirinae* on one side (VP1-VP4-VP2) and the remaining *Microviridae* subgroups on the other (VP1-VP2-VP4). The finding that microphages belonging to different subgroups occupy a variety of different ecological niches suggests that the association of microphages with bacteria is ancient, possibly predating the divergence of this cellular domain into the contemporary lineages.

Major Capsid Protein Structure and Evolution

Based on similarity in virion architecture, it has been previously suggested that members of the *Microviridae* might share a common origin with eukaryotic viruses from the families *Circoviridae*, *Geminiviridae* and *Parvoviridae* [10]. Indeed, all these ssDNA viruses utilize eight-stranded β-barrel capsid proteins to build their icosahedral (T = 1) virions [36,37]. However, while the capsid proteins of *Microviridae* and parvoviruses possess long insertion loops connecting the β-strands (although at different locations; [36]), those of geminiviruses and circoviruses are much more compact [37,38]. Consequently, if the structural relationship indeed testifies for the common origin of these viruses, the evolution of *Microviridae* virion structure most likely proceeded through acquisition of insertion loops within the eight-stranded β-barrel core. As revealed through comparative analysis and structural modeling of the MCPs presented in this study, such dynamics within the loop regions of microviral MCPs appears to be an ongoing process, possibly assisting host-range expansion and adaptation to new environments in this viral family. This is especially obvious for microphages associated with human microbiota (all alpaviruses and certain gokushoviruses) that on average possess larger and more numerous insertions within their MCPs (Fig. 1, Fig. S4B). Paradoxically, although ΦX174-like phages are also known to infect hosts isolated from human samples [39], their MCPs are the most compact among the *Microviridae*. Interestingly, the putative receptor-binding spikes present at the three-fold symmetry axes of gokushovirus capsids ([21]; Fig. 2A) are also likely to decorate the virions of alpaviruses [17] and pichoviruses (Fig. 2C). The presence of this protrusion in all gokusho-, alpa- and pichoviruses, suggests that this feature is ancestral to the spikes present at the five-fold vertices of ΦX174-like microvirus capsids.

The number and the size of insertions within the microviral MCPs were similar in both prophages and free-living viruses, suggesting that these sequence modifications do not preclude the formation of viable virions. This specific evolutionary pattern of human microbiome *Microviridae* MCPs is reminiscent to co-evolution consequences described for cultivated phages. As described in the experiment of Paterson *et al.* [40], the basis of co-evolution is the absence of "non-adapted" host for the phage. This is consistent with the restriction of these viruses to human gut flora, where the highest bacterial densities for a microbial habitat were found [41]. Thus, human gut *Microviridae* are likely to be exposed to a higher host-phage encounter frequency compared to other *Microviridae*, thereby increasing the evolution rate of their MCP.

Horizontal Gene Acquisition of a Possible Endolysin Gene

Until now, genes encoding novel functions in microviral genomes were thought to emerge from pre-existing genomic regions through accumulation of point mutations [19,33]. However, the discovery in the *Microviridae* genomes of peptidase coding genes that were clearly acquired by horizontal gene transfer (HGT), and more likely through at least two independent transfer events, shows that *Microviridae* are able to integrate genes of interest from external sources into their genomes, even if such transfers are rare. The different uncharacterized ORFs detected in the new *Microviridae* genomes are then of great interest, since they could represent other horizontally acquired genes. On the contrary, no direct gene transfers between two *Microviridae* genomes could be detected in our dataset. Notably, phylogenetic analysis of the 47 closely-related *Escherichia coli*-infecting microviruses illuminated a few cases of HGT between these viruses that probably occurred by homologous recombination [14]. It is possible that HGT in *Microviridae* is limited by the genetic distance between the donor and the recipient virus species. Consequently, larger datasets of closely related virus genomes might be needed to better understand the prevalence of homologous recombination-driven HGT events in *Microviridae*.

ΦX174 is the only microvirus for which the mechanism of host cell lysis has been elucidated. Unlike dsDNA phages that typically encode a holin-endolysin system, where holin perforates the cytoplasmic membrane and endolysin digests the peptidoglycan, microviruses depend on a single-gene lysis system [42]. It has been shown that protein E of ΦX174 induces lysis by inhibiting cell wall biosynthesis [43]. It was therefore surprising to discover that gene for M15_3 peptidase identified in several gokushoviral and alpaviral genomes is associated with canonical lysis genes (for holin and endolysin) in dsDNA (pro)phages (Fig. 3A), suggesting that phage-encoded M15 peptidases might play a role in cell lysis during virus progeny release. Indeed, endopeptidase PLY500 (family VanY; PF02557), which is structurally related to M15 family proteases (families M15_3 and VanY belong to the same clan – Peptidase_MD; CL0170), acts as an endolysin at the end of the infection cycle of *Listeria* phage A500 [44]. We therefore suggest that *Microviridae* M15 peptidase might also be involved in dissolution of the host cell wall at the end of the phage life cycle.

Microviridae Life Cycle and Putative Bacterial Hosts

From the current knowledge on *Microviridae*, only one subgroup (*Alpavirinae*) was found to contain temperate members (i.e. detected as prophages). This could be linked to a relatively low number of complete bacterial genomes from aquatic environments. However, the absence of *Microviridae* prophages in *Enterobacteria*, which have been far more thoroughly studied, as well as the absence of detection of any new prophage even with the new *Microviridae* genomes described here, suggests that the use of the lysogenic cycle is likely to be rare among *Microviridae*.

The lysogenic cycle of some of the *Alpavirinae* and the presence of a horizontally transferred peptidase in several of their genomes made it possible to deduce potential host organisms for these phages. As *Alpavirinae* prophages have been found only within *Bacteroidetes* genomes, and the *Alpavirinae* peptidase genes are most similar to genes from *Bacteroidetes* genomes as well, members of the *Alpavirinae* group are likely to infect members of this bacterial phylum. Interestingly, free-living *Alpavirinae* closely-related to *Prevotella* prophages were only found in a lung sample, whereas *Alpavirinae* related to *Bacteroides* prophages were found in different human stool samples. This finding is consistent with the fact that most of the sequenced *Prevotella* strains have been isolated from oral samples, while *Bacteroides* are thought to be primarily associated with gut flora.

The absence of described prophage for *Gokushovirinae* makes it impossible to be conclusive regarding the potential host(s) of these viruses. Yet, 4 peptidases from human gut gokushoviruses form a common clade with genes from a marine bacterium (*Ahrensia* sp. R2A130) and 7 human gut bacteria belonging to genera *Faecalibacterium* and *Anaerofustis* (Fig. 3B, Fig. S9). Both of these genera are members of the order *Clostridiales* (phylum *Firmicutes*), thus it is tempting to speculate that at least some members of the *Gokushovirinae* might infect Gram-positive bacteria.

Materials and Methods

Viromes Data Set

A set of 95 viromes available in public databases were downloaded and used in this study (Table S1). Lake Pavin and Lake Bourget viromes were previously described in [45], viromes identified with a number from 12 to 87 in [46], Lake Limnopolar in [6], human lung viromes in [47], human gut and human faeces viromes in [25] and [18], hot springs viromes in [48] and virome JCVI_mv858 is part of the GOS dataset [49]. The 95 viromes span viral communities from the 3 main aquatic ecosystems studied so far (i.e. seawater, freshwater and hypersaline) as well as communities associated with different eukaryotes (fish, coral and mosquito), human lungs and human gut.

Complete Genome Identification

All viromes were assembled (Table S1), and screened for circular contigs with significant sequence similarity with *Microviridae* genes (tBLASTx, threshold of 50 on bit score). Viromes were assembled using Newbler 2.6 (454 Life Sciences), using the stringent threshold of 98% identity on 35 bp. In addition to the contigs assembly, Newbler software detect putative links between different contigs, usually used to create scaffolds. The contigs linked to themselves (i.e. the end of the contig is similar to the start of it) were thus considered as circular DNA sequences, and searched for *Microviridae*-like genes via tBLASTx (threshold of 50 on bit score). After a first iteration of this search step, all *Microviridae* genomes retrieved were used as query in a second iteration, to detect more distant homologies (i.e. contigs with genes not significantly similar to known *Microviridae*, but significantly similar to contigs retrieved in the first iteration). A fasta file containing the raw sequences of the 81 *Microviridae* genomes assembled in this study, alongside the annotation of each genome in separated genbank-formatted files, in a zip archive (available through Dryad Digital Repository, doi:10.5061/dryad.8ht80; http://dx.doi.org/10.5061/dryad.8ht80).

In addition, complete bacterial genomes from Refseq database and genomes currently assembled from the NCBI were looked for *Microviridae*-like genes via tBLASTx (threshold of 50 on bit score) in order to identify new prophages related to *Microviridae*.

Annotation of Complete Genomes

An ORF prediction was processed using Glimmer 3.02 [50] for each circular contig identified as a *Microviridae* genome. The predicted ORFs were compared to the sequence database NR using BLASTp and best BLAST hit were conserved. In order to identify and annotate genes not predicted by Glimmer, intergenic regions of the genomes were also compared to NR using BLASTx.

Analysis of Proteins from the New Circular Genomes

Sequences similar to the Major Capsid Protein (VP1 in *Gokushovirinae*, Protein F in *Microvirus*) were retrieved from known sequenced genomes and complete genomes generated from viromes. The VP1 sequence retrieved as prophage in *Prevotella bergensis* genome was not included in the analysis, since the prophage is split among two scaffolds [17], and thus a genome map is difficult to draw from it. Still, its presence did not modify the tree topology.

A multiple alignment of these VP1 protein sequences was done using Muscle [51]. Mega 5 [52] was used to generate a Neighbor-joining phylogenetic tree from this alignment. A custom-designed Perl script was used to calculate the percentage of identity between each pair of protein sequences, based on the multiple alignments computed with Muscle [51]. Jalview [53] was used to visualize RCR I motif manually on the multiple alignment.

Structural Modeling and Model Quality Assessment

VP1 homologues from each of the analyzed *Microviridae* subgroup were aligned using PROMALS3D [54] and analyzed for the presence and location of insertions with respect to the sequence of ΦX174 protein F [11]. VP1 sequences of Pavin_279 and BMV5 prophage from *Prevotella bucalis* (GI:282877220; [17]) were chosen as representatives of subgroups *Pichovirinae* and *Alpavirinae*, respectively. Three-dimensional model of the Pavin_279 VP1 was generated using a multi-template (ΦX174 F, PDB ID:1CD3 and SpV4 VP1, PDB ID:1KVP) modeling with MODELLER v9.10 [55] The BMV5 VP1 model was obtained using I-TASSER, which uses a combination of *ab initio* and homology-based approaches for structural modelling [56]. The initial Pavin_279 and BMV5 VP1 models were optimized via multiple rounds of loop refinement with MODELLER v9.10. The stereochemical quality of the models was then assessed with ProSA-web [57]. The final Pavin_279 and BMV5 VP1 models had the quality Z-scores of −6.57 and −6.73, respectively, which were comparable to those of the template structures (-6.4 for ΦX174 F and −6.14 for SpV4 VP1). Comparison and visualization of the structural models was performed with VMD [58] and UCSF Chimera [59].

Peptidase Phylogenetic Tree

Reference peptidase sequences were taken from the NR database, based on a BLASTp of the *Microviridae* peptidases (threshold of 90 on bit score). Peptidases from *Bacteroides* and

Prevotella genomes were added to the dataset, as *Microviridae* prophages had been detected in each of these genera. The multiple alignment was computed using Muscle [51], and the maximum-likelihood phylogenetic tree was computed with FastTree [60].

Major Capsid Protein Detection and Affiliation from Linear Sequences

The unassembled reads from the set of viromes used in this study were screened for sequences homologous to major capsid protein, and these sequences were affiliated via a best BLASTx hit against a database formed of all VP1 from the complete *Microviridae* genomes both published and assembled in this study (threshold of 50 on BLAST bit score).

Supporting Information

Figure S1 Boxplot of genome sizes within each clade. Affiliations were based on the major capsid protein phylogenetic tree (Fig. 1).
(TIFF)

Figure S2 Heatmap based on the percentage of identity computed from the major capsid protein multiple alignment. Scale is indicated on the top left, with the distribution of the percentages of identity. The genome affiliation is indicated above the map, and groups are framed on the heatmap.
(TIFF)

Figure S3 Multiple amino acid alignment of the major capsid protein. Large insertions (more than 10 aa) are framed and identified from A to G. The insertion retrieved in all *Microviridae* but *Enterobacteria* phages known to induce mushroom-like structure is identified as the insertion E. One or several sequences were taken for each group, ϕX174 for *Enterobacteria* phages, CF7ML00001 and *Prevotella Buccalis* for *Alpavirinae*, Pavin_00723 for *Pichovirinae*, *Chlamydia* phage Chp2 and Bourget_00154 for *Gokushovirinae* and *Spiroplasma* phage 4.
(TIFF)

Figure S4 A boxplot illustrating length variation of the 'mushroom-like' protrusion-forming insertions in the major capsid proteins of *Gokushovirinae*, *Alpavirinae*, and *Pichovirinae*. The insertion lengths are plotted as a function of the *Microviridae* subgroup (A) and ecosystem type (B).
(TIFF)

Figure S5 Alpaviral VP1 in the context of the entire virion. Pseudoatomic model of the gokushovirus SpV4 virion (PDB ID:1KVP) with one of the capsomers substituted with the structural model of the alpaviral VP1 (*Prevotella bucalis* prophage BMV5). The hot-spots in the alpaviral VP1s where specific insertions (>15 aa) with respect to the BMV5 VP1 sequence were detected are indicated with orange spheres. The length of the largest insertion at each of the hot-spots is indicated along with the name of a corresponding viral genome. HF, human feces.
(TIFF)

Figure S6 Alignment of the conserved motifs of the superfamily I rolling-circle replication protein.
(TIFF)

Figure S7 Heatmap based on the percentage of identity from the replication protein multiple alignment. Scale is indicated on the top left, with the distribution of the percentages of identity. The genome affiliation is indicated above the heatmap, and groups are framed on the heatmap.
(TIFF)

Figure S8 Heatmap based on the percentage of identity detected on the capsid assembly protein multiple alignment. Scale is indicated on the top left, with the distribution of the percentages of identity. The genome affiliation is indicated above the heatmap, and groups are framed on the heatmap.
(TIFF)

Figure S9 Maximum-likelihood phylogenetic tree based on peptidase M15_3 protein sequences. Each reference sequences is identified by its name, followed by its gene id. *Alpavirinae* sequences are highlighted in pink, *Gokushovirinae* in blue, and viral reference sequences are marked with a black circle.
(TIFF)

Table S1 List of viromes assembled. For each virome, the number of circular contigs identified as complete *Microviridae* genome is indicated. The web-server hosting the datasets are : NCBI (www.ncbi.nlm.nih.gov), MG-Rast (http://metagenomics.anl.gov), and Metavir (http://metavir-meb.univ-bpclermont.fr). When available, the methodology used to purify viral particle is indicated (CsCl : Cesium Chloride, PEG : Polyethylene Glycol, LASL : linker amplified shotgun library and MDA : phi29-mediated multiple displacement amplification). *2 contigs were detected for virome 35 Marine_Sar_Vir, but they corresponded to the 2 contigs already assembled from this virome, described in Tucker et al., 2011, and were thus discarded.
(DOC)

Table S2 List of circular contigs similar to complete genomes of *Microviridae*. For each major protein, the gi of the best BLAST hit is indicated with the bit score of the corresponding BLAST. All the sequences and corresponding annotations are available through Dryad Digital Repository, doi:10.5061/dryad.8ht80; http://dx.doi.org/10.5061/dryad.8ht80.
(DOC)

Table S3 List of the *Microviridae* peptidase genes detected, with their best BLAST hit against NR database.
(DOC)

Author Contributions

Conceived and designed the experiments: SR DD FE. Performed the experiments: SR AP. Analyzed the data: SR MK FE. Contributed reagents/materials/analysis tools: SR MK. Wrote the paper: SR MK DD FE.

References

1. Wommack KE, Colwell RR (2000) Virioplankton: viruses in aquatic ecosystems. Microbiol Mol Biol Rev 64: 69–114.
2. Hatfull GF, Hendrix RW (2011) Bacteriophages and their genomes. Curr Opin Virol 1: 298–303.
3. Miller RV (2001) Environmental bacteriophage-host interactions: factors contribution to natural transduction. Antonie Van Leeuwenhoek 79: 141–147.
4. Hendrix RW (2002) Bacteriophages: evolution of the majority. Theor Popul Biol 61: 471–480.
5. Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, et al. (2009) The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. PLoS Comput Biol 5: e1000593.
6. López-Bueno A, Tamames J, Velázquez D, Moya A, Quesada A, et al. (2009) High diversity of the viral community from an antarctic lake. Science 326: 858–861.
7. Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F (2007) Metagenomic analysis of the microbial community associated with the coral porites astreoides Environ Microbiol 9: 2707–2719.

8. Rosario K, Duffy S, Breitbart M (2009) Diverse Circovirus-like genome architectures revealed by environmental metagenomics. J Gen Virol 90: 2418–2424.
9. Tucker KP, Parsons R, Symonds EM, Breitbart M (2011) Diversity and distribution of single-stranded dna phages in the north atlantic ocean ISME J 5: 822–830.
10. Cherwa JE, Fane BA (2011) *Microviridae*: microviruses and gokushoviruses. In: Encyclopedia of life sciences. Ltd., Chichester, United Kingdom. doi:10.1002/9780470015902.a0000781.pub2.
11. McKenna R, Xia D, Willingmann P, Ilag LL, Krishnaswamy S, et al. (1992) Atomic structure of single-stranded dna bacteriophage phix174 and its functional implications. Nature 355: 137–143.
12. Bernal RA, Hafenstein S, Esmeralda R, Fane BA, Rossmann MG (2004) The phix174 protein J mediates dna packaging and viral attachment to host cells. J Mol Biol 337: 1109–1122.
13. Morais MC, Fisher M, Kanamaru S, Przybyla L, Burgner J, et al. (2004) Conformational switching by the scaffolding protein D directs the assembly of bacteriophage phix174. Mol Cell 15: 991–997.
14. Rokyta DR, Burch CL, Caudle SB, Wichman HA (2006) Horizontal gene transfer and the evolution of microvirid coliphage genomes. J Bacteriol 188: 1134–1142.
15. Lee HS, Sobsey MD (2011) Survival of prototype strains of somatic coliphage families in environmental waters and when exposed to uv low-pressure monochromatic radiation or heat. Water Res 45: 3723–3734.
16. Carstens EB (2010) Ratification vote on taxonomic proposals to the international committee on taxonomy of viruses (2009). Arch Virol 155: 133–146.
17. Krupovic M, Forterre P (2011) *Microviridae* goes temperate: microvirus-related proviruses reside in the genomes of *Bacteroidetes*. PLoS One 6: e19893.
18. Kim M, Park E, Roh SW, Bae J (2011) Diversity and abundance of single-stranded dna viruses in human feces. Appl Environ Microbiol 77: 8062–8070.
19. Fane BA, Brentlinger KL, Burch AD, Chen M, Hafenstein S, et al. (2011) the *Microviridae*. In: The bacteriophages. Oxford Press. 129–145.
20. Brentlinger KL, Hafenstein S, Novak CR, Fane BA, Borgon R, et al. (2002) *Microviridae*, a family divided: isolation, characterization, and genome sequence of phiMH2k, a bacteriophage of the obligate intracellular parasitic bacterium bdellovibrio bacteriovorus. J Bacteriol 184: 1089–1094.
21. Chipman PR, Agbandje-McKenna M, Renaudin J, Baker TS, McKenna R (1998) Structural analysis of the *Spiroplasma* virus, spv4: implications for evolutionary variation to obtain host diversity among the *Microviridae*. Structure 6: 135–145.
22. Renaudin J, Pascarel MC, Bové JM (1987) Spiroplasma virus 4: nucleotide sequence of the viral dna, regulatory signals, and proposed genome organization J Bacteriol 169: 4950–4961.
23. Zsak L, Day JM, Oakley BB, Seal BS (2011) The complete genome sequence and genetic analysis of ϕca82 a novel uncultured microphage from the turkey gastrointestinal system. Virol J 8: 331.
24. Desnues C, Rodriguez-Brito B, Rayhawk S, Kelley S, Tran T, et al. (2008) Biodiversity and biogeography of phages in modern stromatolites and thrombolites. Nature 452: 340–343.
25. Minot S, Sinha R, Chen J, Li H, Keilbaugh SA, et al. (2011) The human gut virome: inter-individual variation and dynamic response to diet. Genome Res 21: 1616–1625.
26. Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, et al. (2006) The marine viromes of four oceanic regions. PLoS Biol 4: e368.
27. Nelson KE, Weinstock GM, Highlander SK, Worley KC, Creasy HH, et al. (2010) A catalog of reference genomes from the human microbiome. Science 328: 994–999.
28. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A (2005) Ffas03: a server for profile–profile sequence alignments. Nucleic Acids Res 33: W284–8.
29. Chapman MS, Liljas L (2003) Structural folds of viral proteins. Adv Protein Chem 64: 125–196.
30. Azuma J, Morita J, Komano T (1980) Process of attachment of phix174 parental dna to the host cell membrane. J Biochem 88: 525–532.
31. Cherwa JEJ, Young LN, Fane BA (2011) Uncoupling the functions of a multifunctional protein: the isolation of a dna pilot protein mutant that affects particle morphogenesis. Virology 411: 9–14.
32. Ruboyianes MV, Chen M, Dubrava MS, Cherwa JEJ, Fane BA (2009) The expression of n-terminal deletion dna pilot proteins inhibits the early stages of phix174 replication. J Virol 83: 9952–9956.
33. Krupovic M, Prangishvili D, Hendrix RW, Bamford DH (2011) Genomics of bacterial and archaeal viruses: dynamics within the prokaryotic virosphere. Microbiol Mol Biol Rev 75: 610–635.
34. Bochtler M, Odintsov SG, Marcyjaniak M, Sabala I (2004) Similar active sites in lysostaphins and d-ala-d-ala metallopeptidases. Protein Sci 13: 854–861.
35. Kim K, Bae J (2011) Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded dna viruses. Appl Environ Microbiol 77: 7663–7668.
36. Bennett A, McKenna T, Agbandje-McKenna M (2008) A comparative analysis of the structural architecture of ssdna viruses. Computational and Mathematical Methods in Medicine 9–34: 183–196.
37. Khayat R, Brunn N, Speir JA, Hardham JM, Ankenbauer RG, et al. (2011) The 2.3-angstrom structure of Porcine Circovirus 2. J Virol 85: 7856–7862.
38. Krupovic M, Ravantti JJ, Bamford DH (2009) Geminiviruses: a tale of a plasmid becoming a virus. BMC Evol Biol 9: 112.
39. Michel A, Clermont O, Denamur E, Tenaillon O (2010) Bacteriophage phix174's ecological niche and the flexibility of its escherichia coli lipopolysaccharide receptor. Appl Environ Microbiol 76: 7310–7313.
40. Paterson S, Vogwill T, Buckling A, Benmayor R, Spiers AJ, et al. (2010) Antagonistic coevolution accelerates molecular evolution. Nature 464: 275–278.
41. Ley RE, Peterson DA, Gordon JI (2006) Ecological and evolutionary forces shaping microbial diversity in the human intestine. Cell 124: 837–848.
42. Young R, Wang IN (2006) Phage lysis. In: The bacteriophages. Oxford Press. p. 104–125.
43. Bernhardt TG, Roof WD, Young R (2000) Genetic evidence that the bacteriophage phi x174 lysis protein inhibits cell wall synthesis. Proc Natl Acad Sci U S A 97: 4297–4302.
44. Korndörfer IP, Kanitz A, Danzer J, Zimmer M, Loessner MJ, et al. (2008) Structural analysis of the l-alanoyl-d-glutamate endopeptidase domain of listeria bacteriophage endolysin ply500 reveals a new member of the las peptidase family. Acta Crystallogr D Biol Crystallogr 64: 644–650.
45. Roux S, Enault F, Robin A, Ravet V, Personnic S, et al. (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. PLoS One 7: e33641.
46. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, et al. (2008) Functional metagenomic profiling of nine biomes. Nature 452: 629–632.
47. Willner D, Furlan M, Haynes M, Schmieder R, Angly FE, et al. (2009) Metagenomic analysis of respiratory tract dna viral communities in cystic fibrosis and non-cystic fibrosis individuals. PLoS One 4: e7370.
48. Schoenfeld T, Patterson M, Richardson PM, Wommack KE, Young M, et al. (2008) Assembly of viral metagenomes from yellowstone hot springs. Appl Environ Microbiol 74: 4164–4174.
49. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, et al. (2007) The sorcerer II global ocean sampling expedition: northwest atlantic through eastern tropical pacific. PLoS Biol 5: e77.
50. Delcher AL, Bratke KA, Powers EC, Salzberg SL (2007) Identifying bacterial genes and endosymbiont dna with glimmer. Bioinformatics 23: 673–679.
51. Edgar RC (2004) Muscle: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.
52. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) Mega5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731–2739.
53. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.
54. Pei J, Grishin NV (2007) Promals: towards accurate multiple sequence alignments of distantly related proteins. Bioinformatics 23: 802–808.
55. Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, et al. (2000) Comparative protein structure modeling of genes and genomes. Annu Rev Biophys Biomol Struct 29: 291–325.
56. Zhang Y (2008) I-tasser server for protein 3d structure prediction BMC Bioinformatics 9: 40.
57. Wiederstein M, Sippl MJ (2007) Prosa-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. Nucleic Acids Res 35: W407–10.
58. Humphrey W, Dalke A, Schulten K (1996) Vmd: visual molecular dynamics. J Mol Graph 14: 33–8, 27–8.
59. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, et al. (2004) Ucsf chimera–a visualization system for exploratory research and analysis. J Comput Chem 25: 1605–1612.
60. Price MN, Dehal PS, Arkin AP (2010) Fasttree 2–approximately maximum-likelihood trees for large alignments. PLoS One 5: e9490.

Diversité et complexité insoupçonnées au sein des Microviridae

L'assemblage de ces viromes publiés a ainsi permis de générer 81 nouveaux génomes de *Microviridae,* issus de différents types d'écosystèmes. L'analyse de ces nouvelles séquences génomiques a permis de décrire un nouveau sous-groupe au sein des *Microviridae*, que nous proposons de nommer "*Pichovirinae"* (pour *pichòt* : petit en occitan). Des différences au niveau de la distributions des différents sous-groupes ont également été notées : si le sous-groupe des *Gokushovirinae* est détecté dans tous les types de milieux, les *Alpavirinae* sont à l'inverse uniquement observés au sein du tractus digestif humain. De

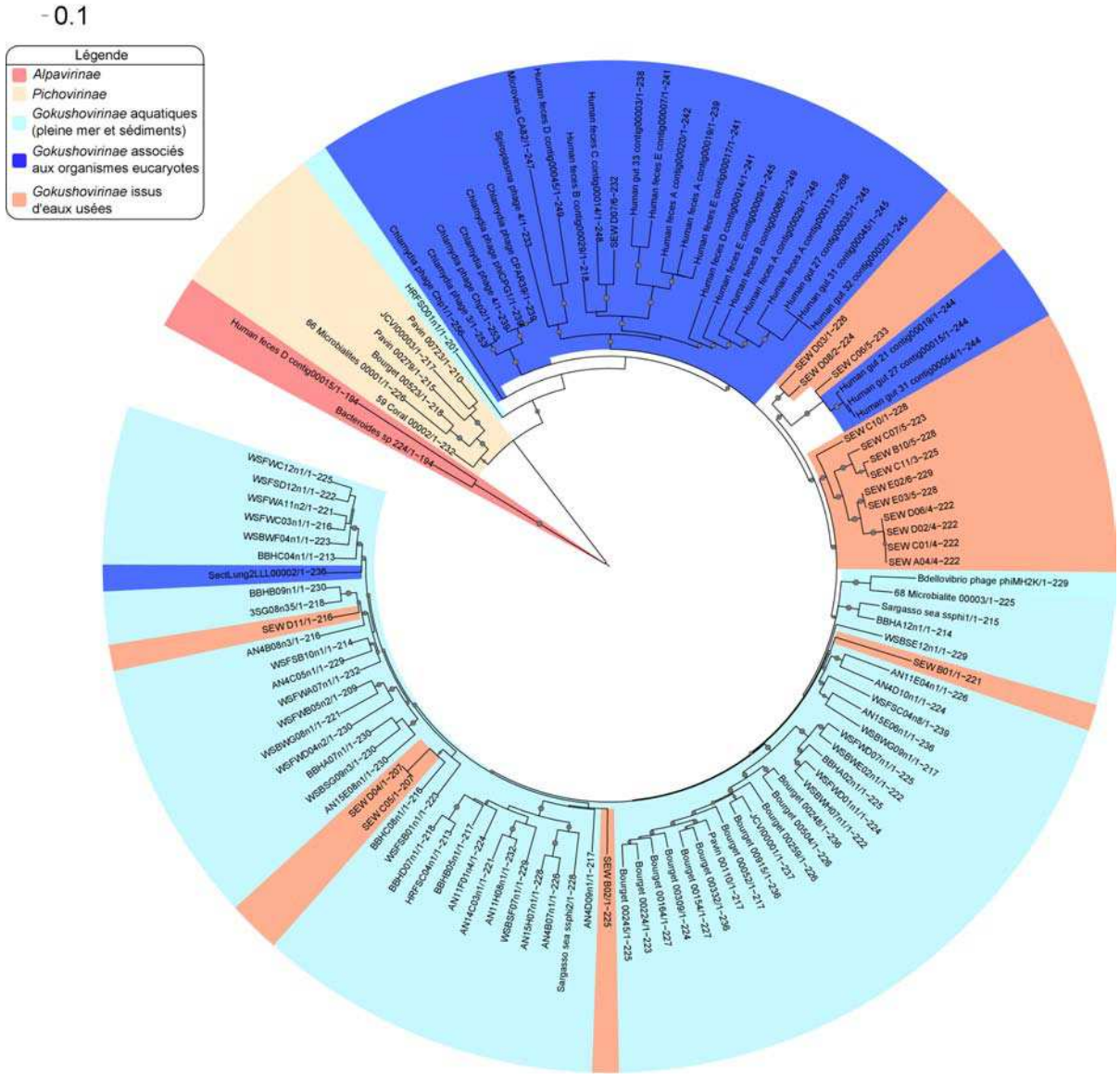


Figure IV.2 : Arbre phylogénétique du sous-groupe des Gokushovirinae basé sur la protéine majeure de capside. *Les sous-groupes des Alpavirinae et Pichovirinae sont utilisés comme groupe externe. Les nœuds dont les scores de bootstraps dépassent 0.5 sont indiqués par un cercle gris.*

manière plus surprenante, aucun Microvirus proche de la souche de référence Enterobacteria phage phiX174 n'a été observé. Ainsi, les souches connues et étudiées en laboratoire ne semblent pas correspondre aux virus effectivement présents en abondance dans l'environnement.

Dans un deuxième temps, la comparaison des génomes a mis en évidence une très forte conservation de l'ordre des gènes, avec un seul événement de recombinaison observé, à l'origine du sous-groupe des *Pichovirinae*. Les génomes assemblés ont également révélé la présence d'un gène codant pour une peptidase d'origine bactérienne au sein de différents génomes de *Microviridae,* visiblement obtenu à la suite de plusieurs événements de transferts indépendants. Il s'agit ainsi du premier transfert de gène décrit pour ces phages avec le génome de l'hôte. Associé à la description récente de prophages de *Microviridae* observés dans des génomes bactériens de la flore intestinale humaine (Krupovic & Forterre, 2011), ces éléments montrent que malgré leur apparente simplicité, ces phages peuvent tout comme les virus à ADN double brin adopter plusieurs types de cycles et agir en tant qu'agent de transfert de gènes.

Enfin, l'analyse des protéines majeures de capsides a également révélé des différences au niveau structurel entre les nouveaux génomes assemblés et les microvirus de type Phix174. En effet, les reliefs externes du virion, codés chez les Microvirus cultivés par un gène spécifique, sont visiblement générés par une large insertion au sein de la séquence du gène codant pour la protéine majeure de capside chez les autres *Microviridae*. Différentes modélisations ont ainsi confirmé que cette insertion, retrouvée sur l'ensemble des nouveaux génomes assemblés, était très variable entre les différents génomes. Cette variabilité est ainsi sans doute associée aux interactions étroites entre la membrane de la cellule hôte et la surface du virion.

Depuis la publication de ces travaux, les *Microviridae* ont à nouveau été détectés dans différents environnements, comme les fosses sous-marines (Yoshida *et al.*, 2013). La distribution de ces virus (plus particulièrement du sous-groupe des *Gokushovirinae*) dans les milieux aquatiques a également été étudiée *via* des approches d'amplification par PCR, d'abord ciblée sur deux génotypes spécifiques (Tucker *et al.*, 2011), puis sur l'ensemble du sous-groupe (Figure IV.2, Hopkins, Breitbart, *et al.,* communication personnelle). Plusieurs types d'échantillons aquatiques ont ainsi été étudiés (océans, lacs, et eaux usées issues de stations d'épuration). Ces résultats semblent indiquer que les *Gokushovirinae* détectés dans les différents milieux aquatiques naturels (lacustres ou marins) appartiennent à un sous-groupe monophylétique, tandis que ceux détectés au sein des eaux usées sont plus hétérogènes. Ces données permettent d'envisager l'hypothèse d'une colonisation des milieux aquatiques naturels par des *Gokushovirinae* initialement associés aux organismes eucaryotes. Ainsi, par un

dialogue entre les différentes méthodologies moléculaires, il est peu à peu possible de reconstituer l'histoire évolutive de ces petits virus sans passer par une étape d'isolement et de culture.

Les virus chimères : un génotype à la croisée des mondes ADN et ARN

Si les *Microviridae* constituent un exemple de famille de petits virus infectant les bactéries, la majeure partie des virus à ADN simple brin décrits infectent des cellules eucaryotes, principalement des animaux et des végétaux. Ces virus possèdent des génomes très réduits (pas plus de quelques kilobases), souvent circulaires, et organisés autour de deux gènes principaux : l'un lié à la réplication et l'autre associé à la capside. Trois familles de petits virus à ADN simple brin, les *Circoviridae*, *Nanoviridae* et *Geminiviridae* partagent ainsi un gène de réplication commun : "Rolling-circle replication associated gene", ou gène

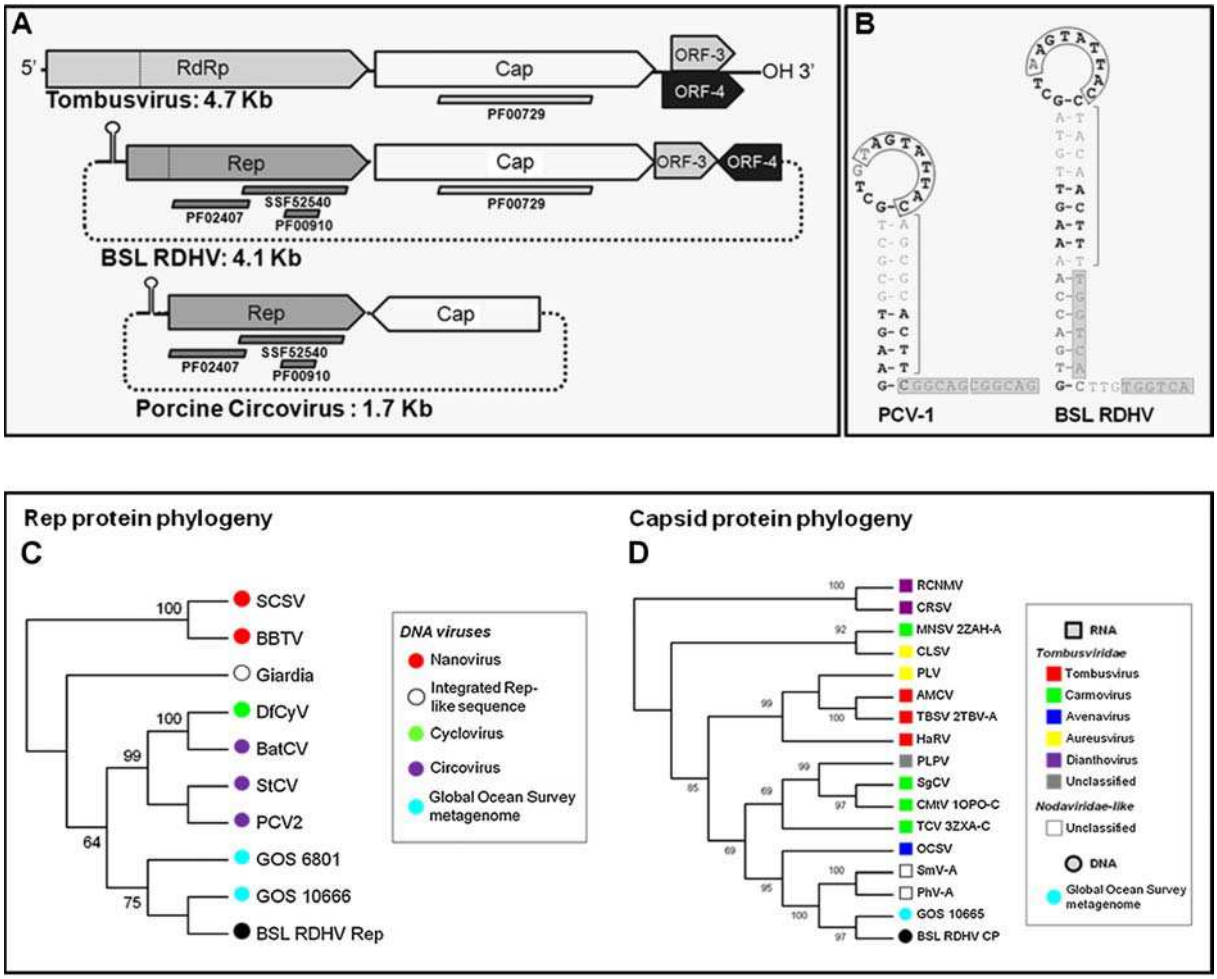


Figure IV.3 : Caractéristiques du premier génome viral chimère ADN-ARN (adapté de Diemer & Stedman, 2012). *A : Carte du génome du virus chimère BSLRDHV et de deux références. B : Structure en épingle à cheveux utilisée pour l'initiation de la réplication chez un Circoviridae, et chez BSLRDHV. C : Arbre phylogénétique issu de l'analyse du gène de réplication. D : Arbre issu de l'analyse du gène de capside.*

associé à la réplication par cercle roulant (RC-Rep). Ces trois groupes ont été particulièrement étudiés, notamment par des approches d'isolement (Sharman *et al.*, 2008; Ilyas *et al.*, 2009) et d'amplification PCR (Ouardani *et al.*, 1999; Mansoor *et al.*, 2005), et différents sous-groupes ont été décrits à partir d'analyses phylogénétiques basées sur ce gène codant pour la protéine de réplication (RC-Rep), identifié comme gène le plus conservé.

En août 2012, Diemer et Stedman ont publié les résultats surprenants d'une analyse de virome issu d'un milieu extrême (lac hyperacide et à forte température), au cours de laquelle ils ont pu observer pour la première fois l'existence d'un petit virus à ADN simple brin visiblement issu de la recombinaison entre deux génomes viraux, l'un à ARN simple brin, et l'autre à ADN simple brin (Diemer & Stedman, 2012). Plus spécifiquement, le génome assemblé possède un gène codant pour une protéine majeure de capside homologue à un gène de *Tombusviridae*, famille de virus à ARN simple brin infectant des plantes et champignons, et un deuxième gène similaire au gène de réplication des *Circoviridae, Nanoviridae* et *Geminiviridae* (Figure IV.3). Les analyses phylogénétiques sur chacun de ces deux gènes semblaient faire état d'un transfert relativement récent du gène codant pour la protéine de capside d'un virus à ARN vers un virus à ADN (Figure IV.3).

Pour mieux comprendre l'origine de ces virus chimères, et déterminer à la fois leur diversité évolutive et leur distribution dans les différents types d'écosystème, nous avons fouillé un ensemble de viromes publiés et assemblés à la recherche de séquences correspondant à des génomes de virus chimères. En assemblant de nouveaux génomes complets, une reconstitution plus précise de la trajectoire évolutive de ces virus, notamment des différents événements ayant suivi le transfert horizontal initial, devait être possible.

Article VII

Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses

Simon Roux, [1,2] François Enault, [1,2] Gisèle Bronner [1,2] Daniel Vaulot, [3] Patrick Forterre, [4;5] and Mart Krupovic, [4]

[1]Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes : Génome et Environnement", Clermont-Ferrand, France

[2]CNRS, UMR 6023, Laboratoire "Microorganismes : Génome et Environnement", Aubière, France

[3]UPMC (Paris-06) and CNRS, UMR 7144, Station Biologique, Roscoff, France

[4]Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extremophiles, Paris, France

[5]Laboratoire de Biologie Moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Orsay, France

En révision, **Nature Communications**

Matériel supplémentaire : Annexe A.8

Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses

Simon Roux[a,b], François Enault[a,b], Gisèle Bronner[a,b], Daniel Vaulot[c], Patrick Forterre[d,e], Mart Krupovic[e1]

[a] Clermont Université, Université Blaise Pascal, Laboratoire "Microorganismes: Génome et Environnement", Clermont-Ferrand, France [b] CNRS UMR 6023, LMGE, Aubière, France [c] UPMC (Paris-06) and CNRS, UMR 7144, Station Biologique, Roscoff, France [d] Institut Pasteur, Unité Biologie Moléculaire du Gène chez les Extrêmophiles, Département de Microbiologie, Paris, France [e] Laboratoire de Biologie Moléculaire du Gène chez les Extrêmophiles, Institut de Génétique et Microbiologie, CNRS UMR 8621, Université Paris Sud, Orsay, France

Keywords: ssDNA viruses, virus evolution, viral metagenomics

Abstract

Viruses with single-stranded (ss) DNA genomes infect hosts from all three domains of life and represent a rapidly expanding supergroup of economically, medically, and ecologically important pathogens. Metagenomic studies have uncovered an astonishing genetic diversity of ssDNA viruses in various environments. Most of them encode replication proteins (Reps) related to those of eukaryotic *Circoviridae*, *Geminiviridae* or *Nanoviridae*; however, exact evolutionary relationships among these viruses remain obscure. Recently, a unique chimeric virus genome has been discovered. The latter has apparently emerged in the course of recombination between ssRNA and ssDNA viruses that respectively donated genes for the capsid (CP) and Rep proteins. Here, we report on the assembly of thirteen new chimeric virus genomes recovered from diverse environments. Our results indicate a single event of CP gene capture from an RNA virus in the history of this virus group. Interestingly, the domestication of the CP gene was followed by (and possibly evoked) an unprecedented recurrent replacement of the Rep genes in chimeric viruses with distantly related counterparts from other ssDNA viruses, including circoviruses, geminiviruses and nanoviruses. We suggest that parasitic and symbiotic interactions between unicellular eukaryotes were central for the emergence of these new virus types and that such turbulent evolution was primarily dictated by incongruence between the genes for CP and Rep. Frequent exchange of Rep genes described here blurs the borders between the major groups of eukaryotic ssDNA viruses and suggests that Reps represent an inadequate marker for tracing their evolutionary history.

INTRODUCTION

Single-stranded (ss) DNA viruses represent a rapidly expanding, diverse supergroup of economically, medically and ecologically important pathogens preying on hosts from all three domains of life. They are classified by the International Committee on Taxonomy of Viruses (ICTV) based on genetic and phenotypic observations into eight families—*Anelloviridae*, *Bidnaviridae*, *Circoviridae*, *Geminiviridae*, *Inoviridae*, *Microviridae*, *Nanoviridae* and *Parvoviridae* (1)—whereas some groups still await proper taxonomical assessment (2-5). ssDNA viruses infecting plants (geminiviruses and nanoviruses) and animals (anelloviruses, circoviruses and parvoviruses) were in the spotlight of extensive research for many years due to their direct effect on the wellbeing of humans. Recently, a previously unsuspected facet of the ssDNA viruses as important players in the global ecosystems has come to light; viruses with ssDNA genomes have been repeatedly isolated from diverse environments, including extreme geothermal (2) and hypersaline habitats (3, 6), soil (7), freshwater and marine ecosystems (8-11).

Whereas some bacterial and archaeal ssDNA viruses display filamentous or pleomorphic (variable appearance) virion morphologies (2, 3, 12), all eukaryotic ones (namely *Anelloviridae, Bidnaviridae, Circoviridae, Geminiviridae, Nanoviridae* and *Parvovriridae*) pack their genomes into small icosahedral capsids, constructed from multiple copies of a single (e.g., geminiviruses) or, as in parvoviruses, several nearly identical capsid proteins (CP) (13). In all cases when high-resolution structural information is available, the CPs of ssDNA viruses were found to display a jelly-roll (antiparallel eight-stranded β-barrel) fold (13-15), which is also found in the vast majority of icosahedral positive-sense ssRNA viruses infecting eukaryotic hosts (16). At the sequence level, however, the similarity between the capsid proteins of viruses belonging to different families is not recognizable. Another feature which is common to eukaryotic ssDNA viruses is the mechanism of genome replication; all these viruses are believed to replicate their genomes via the rolling-circle (RC) or catalytically similar rolling-hairpin

mechanism mediated by homologous virus-encoded RC replication initiation proteins (RC-Rep) (17). In this respect, ssDNA viruses resemble prokaryotic RC plasmids, pointing towards a possible evolutionary link between these two types of mobile genetic elements (17, 18). A characteristic feature of RC-Reps of eukaryotic ssDNA viruses is the presence of the superfamily 3 helicase (S3H) domain (19, 20), which is fused C-terminally to the catalytic nuclease domain encompassing the three signature motifs found in all prokaryotic and eukaryotic virus and plasmid RC-Reps (17). As opposed to CPs, RC-Reps of eukaryotic viruses display actual sequence similarity and RC-Rep-based phylogenies recapitulate the major taxonomic groups defined by the ICTV (1, 21, 22). It should be noted, however, that not all eukaryotic ssDNA viruses possess genes for canonical RC-Reps; for example, anelloviruses—even though believed to replicate via RC mechanism—do not encode a protein that would contain the entire set of motifs characteristic to RC-Reps (20).

The origin(s) and evolutionary relationships between ssDNA viruses belonging to different families remain obscure. Structural similarity between the CPs of bacterial microviruses and eukaryotic parvoviruses, circoviruses and geminiviruses (13, 14) was suggested to testify for the common origin of these viruses (23). Alternatively, similarity between the RC-Reps of ssDNA viruses and prokaryotic plasmids on one hand (17, 18, 24) and structural similarity between the CPs of viruses with ssDNA and ssRNA genomes on the other (15), led to the proposal that different groups of ssDNA viruses have emerged from plasmids by acquisition of CP-coding genes from RNA viruses, possibly on multiple independent occasions (16, 24, 25). Indeed, both homologous and illegitimate recombination play important roles in driving the evolution of ssDNA viruses (22).

During the past few years numerous studies on uncultivated viral communities using metagenomic approaches have revealed that genetic diversity of ssDNA viruses is much greater than originally recognized (reviewed in REFS (20, 21)). Many of these uncultivated viruses are related to members of the bacteriophage family *Microviridae* (11), but perhaps even larger number encode RC-Reps displaying phylogenetic affinity to one of three families of eukaryotic ssDNA viruses – *Circoviridae*, *Geminiviridae* and *Nanoviridae* (e.g., (26-28)). Interestingly, instead of encoding genes for corresponding CPs (circo-, gemini- and nano-like), these viruses typically bear open reading frames (ORF) that do not share appreciable similarity with sequences in the databases. Potentially, the lack of recognizable sequence similarity might be caused by the extremely high mutation rates characteristic to ssDNA viruses (29-31). Thus, to navigate in the constantly increasing pool of environmental viral genomes, RC-Reps are often used as markers for classification of the uncultivated ssDNA viruses.

Recently, Diemer and Stedman have described a novel chimeric viral (CHIV) ssDNA genome recovered from a hot, acidic Boiling Springs Lake (BSL), USA (32). Whereas the RC-Rep of the virus was most similar to those of circoviruses, the CP was highly similar to the CPs of ssRNA viruses of the family *Tombusviridae* and two unclassified oomycete-infecting viruses, *Sclerophthora macrospora* virus A (SmV-A) and *Plasmopara halstedii* virus A (PhV-A) (32). Notably, the tombusvirus-like CP topology has not been previously observed for any DNA virus, suggesting that the virus has emerged via recombination between a DNA and an RNA virus. The validity of the assembled viral genome, tentatively named the RNA-DNA hybrid virus (BSL RDHV), and its presence in the lake sediment pore water were confirmed by PCR amplification (32). Importantly, the finding that RNA and DNA viruses recombine to produce novel chimeric entities rationalizes some of the puzzles of the virosphere and allows assessing new hypotheses on the origin and evolution of different viral groups (16). Here, we report on the assembly of thirteen new CHIV genomes recovered from various environments and encoding tombusvirus-like CPs and unexpectedly diverse RC-Reps related to the corresponding proteins of eukaryotic ssDNA viruses belonging to three different families.

RESULTS AND DISCUSSION

New chimeric viruses. To get further insight into the diversity and evolution of CHIVs, we have assembled sequence reads from 103 publicly available viromes and searched the resultant contigs for co-occurrence of genes encoding RC-Reps and RNA virus-like CPs (Table S1). As a result, nine contigs were assembled from viromes derived from atmospheric (33) and aquatic (28, 34, 35) samples. Since ssDNA viruses are known to integrate into the genomes of their hosts (36, 37), we also searched for the presence of CHIVs in the eukaryotic genome databases. The latter approach yielded four additional contigs matching our criteria. Three of these represented contigs from two different whole genome shotgun (WGS) libraries of the photosynthetic picoeukaryote populations dominated by green alga *Bathycoccus* (38), while the fourth one was from the WGS library of *Astrammina rara*, a foraminiferan protist (39). The association of different CHIV contigs with two populations of picoeukaryotes raises an intriguing possibility that unicellular eukaryotes serve as hosts for at least some CHIVs. It is worth noting that foraminiferans often establish an endosymbiotic relationship and were found to host unicellular algae belonging to diverse lineages, including green algae, red algae, diatoms, and dinoflagellates (40). Consequently, it is possible that the CHIV contig associated with *A. rara* derives from a heterologous organism.

General characteristics of the thirteen CHIV genomes (CHIV1–13) obtained by the two approaches are summarized in Table 1. In accordance with the experimentally verified topology of the BSL RDHV genome (32), most (9 out of 13) of the CHIV contigs obtained here were circular. Importantly, the potential stem loops containing nonanucleotide sequences, which serve as origins of replication in ssDNA viruses with circular genomes (20), were readily identifiable in proximity of the RC-Rep genes in all CHIV genomes (Table 1 and Fig. S1). Besides the CP and RC-Rep genes,

	Structure	Length	RC-Rep type	RC-Rep vs Capsid orientation	Number of additional genes	Nonanucleotide + ITR coordinates	Nonanucleotide sequence	Stem loop location	Stem loop strand
Reference ssDNA viruses									
Porcine circovirus 2 / *Circoviridae*	circular	1767	Circo	Reverse	0	10 – 20 | 31 – 40	TAGTATTAC	Before Rep	Same strand
Cyclovirus bat/USA/2009 / Cyclovirus	circular	1703	Circo	Reverse	0	59 – 73 | 85 – 99	TAGTATTAC	Before Rep	Reverse strand
Milk vetch dwarf virus (segment 1) / *Nanoviridae*	circular	1007	Nano	-	0	1 – 11 | 23 – 33	TAGTATTAC	Before Rep	Same strand
Maize streak virus / *Geminiviridae*	circular	2690	Gemini	Reverse	3	2512 - 2529 | 2542 - 2559	TAATATTAC	Before Rep	Reverse strand
Sclerotinia sclerotiorum hypovirulence associated DNA virus 1	circular	2166	Gemini	Reverse	0	2156 - 2163 | 9 - 16	TAATATTAT	Before Rep	Reverse strand
Described chimeric viruses									
BSL RDHV	circular	4100	Circo	Forward	2	1 – 17 | 29 – 45	AAGTATTAC	Before Rep	Same strand
Assembled chimeric viruses									
Seawater virome - 35 Marine contig3 - CHIV1	circular	3802	Circo	Forward	2	5 – 17 | 29 – 41	TAGTATTAC	Before Rep	Same strand
Freshwater virome - Lake Pavin contig15342 - CHIV2	circular	5733	Circo	Reverse	4	1028 – 1037 | 1066 – 1075	CTGTATTAC	After Rep	Same strand
Seawater eukaryote metagenome - Euk T142 contig705 – CHIV3	circular	4675	Circo	Reverse	4	27 - 34 | 68 - 75	TATTATTAC	Before Rep	Same strand
Seawater eukaryote metagenome – Euk T149 contig609 – CHIV4	circular	4677	Circo	Reverse	4	57 – 64 | 98 - 105	TATTATTAC	Before Rep	Same strand
Atmosphere virome - Airborne RD1 contig10 – CHIV5	circular	3354	Circo	Reverse	1	1149 – 1158 | 1170 – 1179	TAGTATTAC	3' End of Rep gene	Reverse strand
Freshwater virome – Lake Bourget contig37546 – CHIV6	circular	3824	Nano	Forward	1	13 -20 | 32 - 39	TAGTATTAC	Before Rep	Same strand
Freshwater virome – Lake Bourget contig37561 – CHIV7	circular	3106	Nano	Forward	0	122 - 132 | 160 - 170	TGTTATTCC	Before Rep	Same strand
Freshwater virome – Lake Pavin contig10824 – CHIV8	linear	3536	Nano	Forward	1	757 - 768 | 799 - 810	AACTATTAT	Before Rep	Reverse strand
Freshwater virome – RW Nursery DNA contig62 – CHIV9	linear	3139	Nano	Forward	0	1575 - 1586 | 1609 - 1620	TAATGTTAC	After Rep	Same strand
Atmosphere virome – Airborne IC2 contig9 – CHIV10	circular	2892	Nano	Reverse	0	1514 – 1526 | 1541 – 1553	GTTTATTAC	After Rep	Reverse strand
Seawater eukaryote metagenome – Euk T149 contig276 – CHIV11	circular	4511	Nano	Reverse	3	94 – 102 | 123 – 131	TATTATTAC	Before Rep	Reverse strand
Foraminifera Whole Genome Sequencing – Astraminna rara contig 97 – CHIV12	linear	3915	Nano	Reverse	2	1616 - 1624 | 1637 - 1645	TAACATTAT	After Rep	Reverse strand
Freshwater virome – Lake Pavin contig403 – CHIV13	linear	3581	Gemini	Reverse	0	268 – 280 | 294 – 306	TAATGTTAT	Before Rep	Reverse strand

Table 1. Characteristics of reference ssDNA viruses genomes, the first chimeric virus (BSL RDHV), and the 13 new chimeric viruses.

some of the CHIVs were predicted to contain up to four additional ORFs. However, sequence analysis did not offer any insight into their possible functions.

Emergence of chimeric viruses is a rare event. All CHIVs were found to encode putative CPs related to those of tombusviruses, to the exclusion of all other groups of RNA viruses. We note that recent exploration of the ssDNA virus diversity associated with dragonflies revealed a viral genome, DfCyclV, encoding a putative protein with weak but significant similarity to the capsid protein of satellite tobacco necrosis virus (STNV) (27). The authors concluded that DfCyclV might be a chimeric virus with a circovirus-like RC-Rep and a tombusvirus-like CP. However, the STNV CP is radically different in sequence and structure from those of tombusviruses and most closely resembles the CPs of geminiviruses (24). Thus, in our opinion, DfCyclV should not be confused with chimeric viruses.

Members of the family *Tombusviridae* have positive-sense ssRNA genomes and infect a variety of land plants (41), although several tombusviruses have been also isolated from freshwater samples (42). Viruses belonging to *Tombusviridae* genera *Aureusvirus*, *Avenavirus*, *Carmovirus*, *Dianthovirus* and *Tombusvirus* possess icosahedral virions with a granular surface. The latter property is determined by a unique domain organization of the CPs of these viruses. Each CP subunit consists of three distinct domains: the N-terminal RNA-binding (R) domain facing the interior of the virion, the shell (S) domain central for the assembly of the icosahedral capsid, and the C-terminal projection (P) domain, which faces away from the capsid surface, giving the virion its granular appearance (Fig. 1A and 1B). Outside of the *Tombusviridae*, the same CP domain organization is expected (based on sequence similarity) only for two recently isolated unclassified oomycete-infecting ssRNA viruses, SmV-A and PhV-A (43, 44).

To better understand the relationship between the CPs of ssRNA viruses and CHIVs, we built a three dimension model of a representative CHIV CP; CHIV10 (Airborne_IC2) was chosen for this purpose (Fig. 1C). In

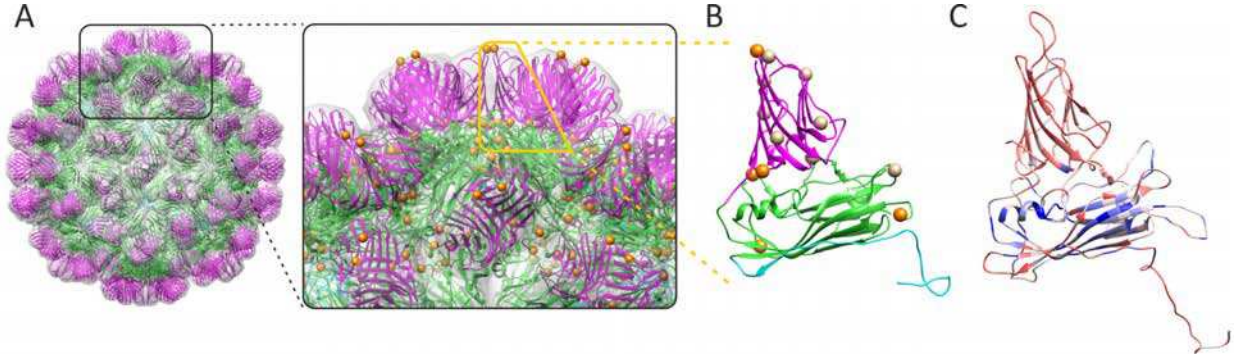


Figure 1. Insights into the structure of chimeric viruses. (*A*) Structure of the melon necrotic spot virus (MNSV; PDB ID:2ZAH) is shown to illustrate the contribution of the distinct capsid protein (CP) domains to the virion organization and the position of these domains in the capsid surface lattice. The projection (P) domain, magenta; shell (S) domain, green; RNA-binding (R) domain, cyan. On the right is a zoom-in on one of the capsid areas, where locations of the conserved insertions present in the CPs of CHIVs as well as viruses SmV-A and PhV-A are indicated with orange spheres. (*B*) Structure of the MNSV CP. The P, S and R domains as well as the locations of conserved insertions are colored as in panel A. In addition, the locations of species-specific sequence insertions present in the CPs of only some CHIVs are indicated with grey spheres. (*C*) Structural model of the CHIV10 CP. The coloring represents sequence conservation among CHIV capsid proteins (red, least conserved; blue, most conserved).

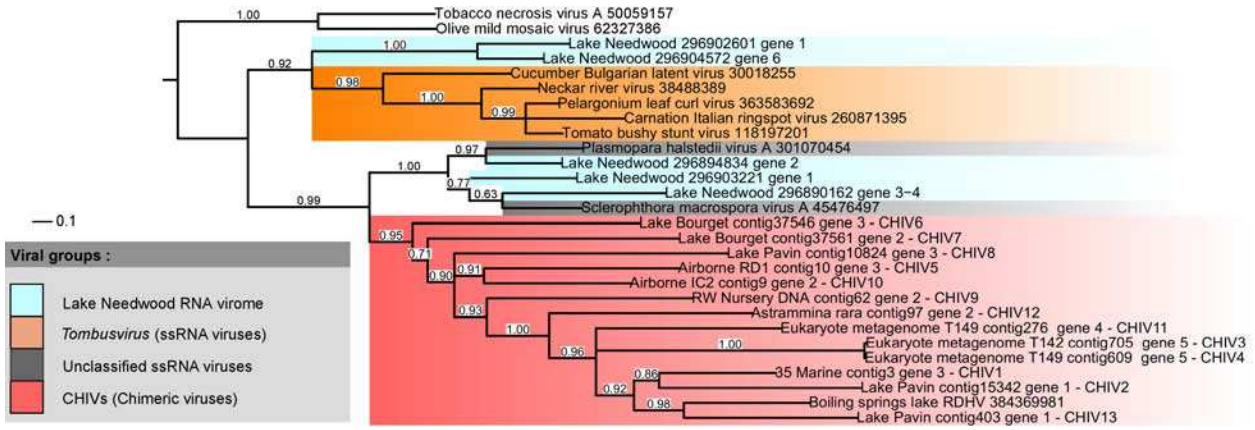


Figure 2. Phylogenetic tree of tombusvirus-like capsid proteins. Chimeric viruses are highlighted in red, tombusviruses in orange, and unclassified ssRNA viruses are either in grey when isolated, or in blue when assembled from Lake Needwood RNA virome. Tobacco necrosis virus A and Olive mild mosaic virus, both members of the genus Necrovirus within Tombusviridae, have CPs lacking the P domain and were used as external group. Bootstrap scores are indicated on internal nodes, and nodes with scores lower than 0.5 were collapsed.

accordance with previous predictions (32), good stereochemical quality of the obtained structural model confirmed that the CPs of CHIVs are likely to display the same structural fold and domain organization as those of tombusviral CPs. Comparison of the fourteen CHIV CPs (13 new and 1 from BSL RDHV) in the context of their tertiary structures revealed that the most conserved part of these proteins corresponds to the S domain, whereas the R and P domains are much more variable (Fig. 1C, Fig. S2). Similar pattern of conservation has been also observed in tombusviruses (45). Closer examination of the multiple alignment of CHIV, tombusviral, oomycete-infecting virus (SmV-A and PhV-A) CP sequences showed that CHIV CPs are more closely related to the proteins of SmV-A and PhV-A than they are to the CPs of tombusviruses. Five unique insertions, not present in tombusviral CPs, are shared between the CPs of CHIVs, SmV-A/PhV-A and the related sequences from the Lake Needwood RNA virome (indicated with orange spheres in Fig. 1A and 1B), which we consider as synapomorphies testifying for the common evolutionary history of these proteins. Furthermore, unlike in tombusviruses, but similarly to SmV-A/PhV-A, CHIV capsids are not likely to be stabilized by calcium ions; none of the CHIV CPs contained the calcium binding motifs, as has been also noted for BSL RDHV (32). Finally, eight species-specific insertions were detected in the CPs of certain CHIVs (grey spheres in Fig. 1B, see also Fig. S2). Most of them were located within the P domains. Importantly, alterations within the P domain are less likely to interfere with capsid formation, which is primarily orchestrated by interactions within the S domain. We hypothesize that P domain is involved in virus-host interaction (possibly host recognition), which would explain its greater variability promoted by a constant arms race between the virus and the host.

To learn on how many independent occasions tombusvirus-like CP genes were captured by DNA viruses, we performed a maximum likelihood phylogenetic analysis of the CHIV, tombusviral and SmV-A/PhV-A CP proteins (Fig. 2). In addition, the dataset was supplemented with tombusvirus-like CP sequences recovered from the RNA virome obtained from Lake Needwood (46). Notably, in none of the datasets, which contained information about both RNA and DNA virus communities present in the same environmental sample (Rosario et al., 2009), could we detect both CHIVs and tombus-like RNA viruses (in the DNA and RNA fractions, respectively), pointing towards their divergent distribution. The tombusvirus sequences formed a well-supported monophyletic clade. Interestingly, all CHIVs clustered together as a sister group to the CPs of SmV-A/PhV-A (Fig. 2). Monophyly of the CHIV CPs and the fact that no other RNA virus-like CPs were found in association with RC-Reps suggest that transfer of a CP gene between RNA and DNA viruses was a unique event and that emergence of chimeric viruses is likely to be rare.

Polyphyly of RC-Reps in chimeric viruses. Sequence analysis of CHIV RC-Reps revealed a domain organization typical of eukaryotic ssDNA viruses, with the N-terminal nuclease domain and the C-terminal S3H domain (17, 20). The three signature motifs of the nuclease domain were readily identifiable in all CHIV RC-Reps, while S3H motifs were conserved in all but two proteins – Walker B motif could not be mapped in the RC-Reps of CHIV6 and CHIV12 (Table S2). Previous analysis of the RC-Rep encoded by BSL RDHV showed that it is most closely related to those of circoviruses (32). Unexpectedly, BLASTp analysis revealed differential affinity of the CHIV RC-Reps to the corresponding proteins from three major groups of eukaryotic ssDNA viruses. The latter observation was subsequently confirmed by phylogenetic analysis of RC-Reps encoded by CHIVs, circoviruses, nanoviruses and geminiviruses (Fig. 3). Like in the case of BSL RDHV, five CHIVs (CHIV1–CHIV5) clustered with circoviruses. CHIV6–12 formed a well-supported phylogenetic clade with nanoviruses, while CHIV13 branched together with geminiviruses, separately from the rest of CHIVs (Fig. 3). Such phylogenetic distribution of CHIV RC-Reps is in stark contrast with the monophyly of the CHIV CPs. Indeed, CHIV pairs which are close on the CP tree fall into different clades on the RC-Rep phylogeny. For example, the three CHIVs recovered from the WGS library of the photosynthetic picoeukaryotes fall into two

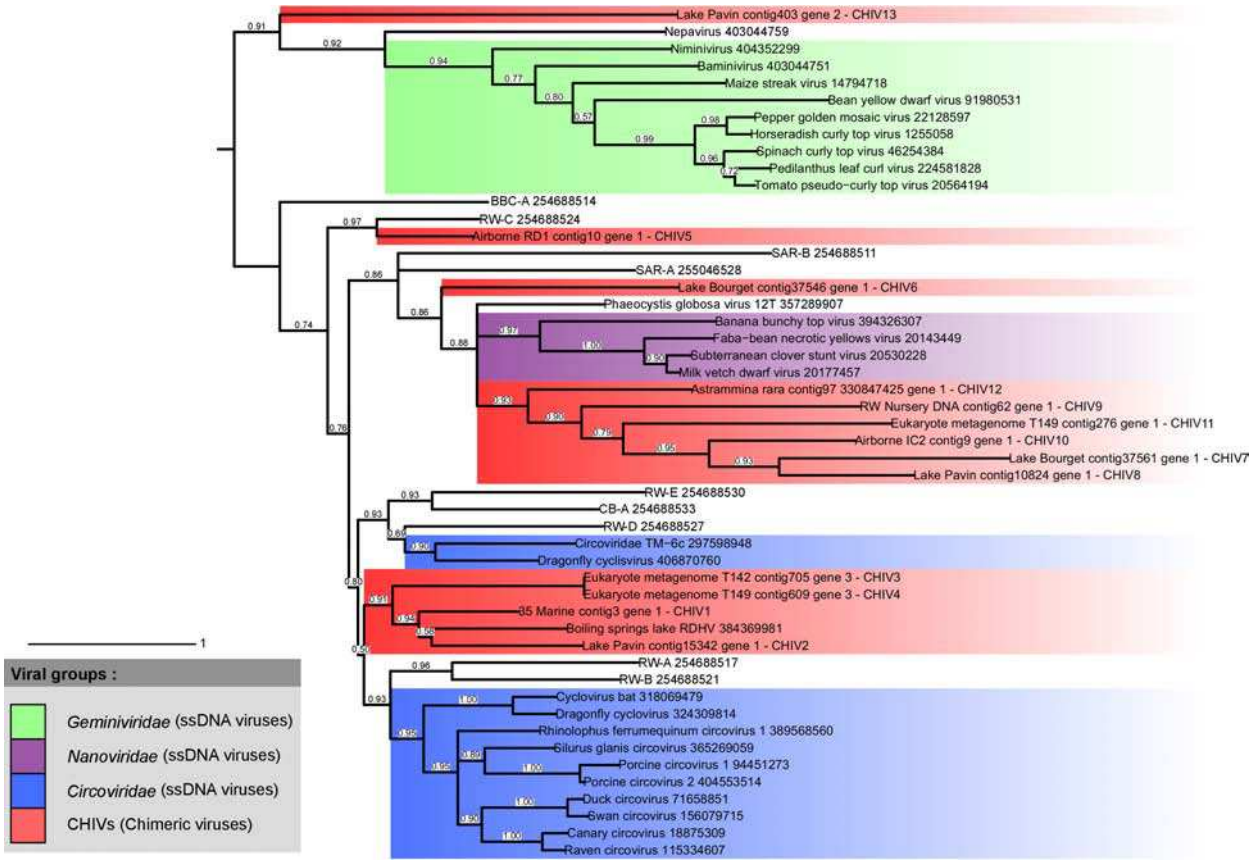


Figure 3. Phylogenetic analysis of the chimeric virus RC-Reps. Chimeric viruses are highlighted in red, Circoviruses in blue, Nanoviruses in purple and Geminiviruses in green. Environmental sequences amplified from Reclaimed Water (RW), Chesapeake Bay (CB) or British Columbia (BBC) were taken as additional references (26, 35), as well as RC-rep gene from a dsDNA algal virus (Phaeocystis globosa virus 12T). All nodes with a bootstrap value lower than 0.5 were collapsed. NCBI GI numbers are indicated for all reference sequences, and bootstraps scores are marked on internal nodes.

different groups (CHIV3 and CHIV4 encode circovirus-like RC-Reps, while CHIV11 has a nanovirus-like protein), despite the fact that their CPs cluster together (Fig. 2). Similarly, the CP of CHIV13 is closely related to the corresponding BSL RDHV protein, but their RC-Reps group with geminiviruses and circoviruses, respectively (Fig. 3).

Recombination is known to play an important role in the evolution of eukaryotic ssDNA viruses (22). However, interfamilial gene exchange has not been convincingly demonstrated for these viruses, suggesting that such recombination might be either uncommon or the recombinants are rarely retained in the population. To compare the evolutionary patterns of CHIVs, circoviruses, nanoviruses, geminiviruses and tombusviruses we have plotted the pairwise distances calculated for CPs from the representative members within each taxon against the corresponding distances between their replication proteins (Fig. 4A). In circoviruses, nanoviruses, geminiviruses and tombusviruses the replication proteins were found to be considerably less divergent that the corresponding CPs. Strikingly, the pattern was found to be the opposite in CHIVs; RC-Reps were much more divergent than in any other virus taxon. In combination with the results of phylogenetic analysis (Fig. 3), such sequence divergence of CHIV RC-Reps is most consistent with multiple independent events of RC-Rep gene replacement in different CHIVs. We hypothesize that the unusually frequent RC-Rep gene transfer in the CHIV lineage could have been instigated by incongruences between the capsid and RC-Rep proteins in the ancestral CHIV. It appears reasonable to assume that CP and RC-Rep, which respectively evolved in the contexts of RNA and DNA viral genomes, would not immediately form a perfect match. Thus, RC-Rep genes could have been exchanged in a search for the optimal CP–Rep combination. However, as soon as the CP and RC-Rep genes are sufficiently adapted to each other (i.e., further 'sampling' decreases fitness) and/or viruses occupy a specific niche where 'sampling' is no longer possible, such high rate of gene exchange is expected to transit to a more conservative mode observed in other eukaryotic ssDNA viruses.

Unicellular algae as recombination hotspots for diverse organisms

Although viromes studied here were assembled from a wide range of biomes (Table S1), CHIVs were exclusively retrieved from aquatic and atmospheric environments. Similarly, when microbial metagenomes were considered, CHIVs once again were identified only in aquatic samples. Three CHIV genomes (two of which are very similar, CHIV3 and CHIV4) were detected in two different samples enriched for the photosynthetic unicellular algae *Bathycoccus*, pointing towards potential association between algae and CHIVs.

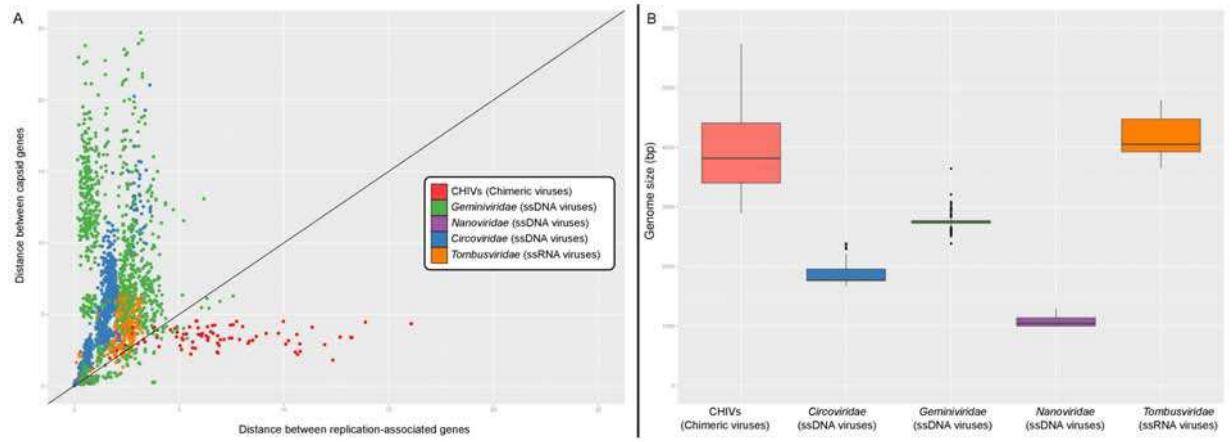


Figure 4. (A) Evolutionary distance (JTT Model) for RC-Rep and capsid protein sequences assessed between pairs of genomes within ssDNA and ssRNA families, and between chimeric viruses. **(B) Boxplot of genome size distribution** in CHIVs, circoviruses, geminiviruses, nanoviruses and tombusviruses.

Interestingly, we identified a close homologue (AET73220; E=4e-29, 35% identity) of CHIV12 RC-Rep (but not the CP) encoded in the genome of a giant dsDNA virus, PgV-12T, infecting *Phaeocystis globosa* (47), photosynthetic unicellular algae. It has been recently demonstrated that satellite viruses and transposons integrate into the genome of the Lentille virus, a relative of mimiviruses (48). It is tempting to speculate that ssDNA viruses and derived elements might represent a new class of molecular parasites preying on giant dsDNA viruses. Regardless, the presence of the RC-Rep gene in the genome of PgV-12T lends additional support to the hypothesis that unicellular algae may host at least some of the CHIVs. More generally, parasitic and symbiotic relationships involving unicellular algae are highly prevalent in aquatic environments (49) and might be central for the emergence of new virus types, such as CHIVs, by providing a unique environment accessible for viruses infecting phylogenetically distant hosts. Such co-localization of various genetic elements of distinct origins and histories could also explain the evolutionary relationships between RC-Reps of prokaryotic plasmids and eukaryotic ssDNA viruses.

Place of chimeric viruses in the virosphere. Metagenomic studies have recently uncovered the unsuspected diversity of ssDNA viruses, many of which encode RC-Reps similar to those of geminiviruses, nanoviruses and, perhaps most commonly, circoviruses (20, 21). However, their CP genes are typically beyond recognition using sequence-based approaches, opening a possibility that these uncultured viruses represent highly divergent yet genuine members of the corresponding viral families. By contrast, CHIVs described here—despite being scattered throughout the RC-Rep phylogeny (Fig. 3)—all share a CP gene, which they apparently inherited from a common ancestor (Fig. 2). Importantly, tombusvirus-like CP gene is not the only feature which distinguishes CHIVs from the three families of eukaryotic viruses mentioned above. CHIV genomes are also significantly larger than those of geminiviruses, nanoviruses and circovirses, and are close in size to the ssRNA genomes of tombusviruses (Fig. 4B). Consequently, capsids larger than those of ssDNA viruses would be required to package such genomes. Interestingly, physical properties, such as the persistence lengths, are similar for ssRNA and ssDNA molecules (50), indicating that tombusvirus-like capsids would be well-fitted to accommodate the larger genomes of CHIVs.

Where do viruses with RNA virus-like capsids, DNA genomes, and RC-Rep diversity spanning the major groups of eukaryotic ssDNA viruses fit in the virosphere? Obviously, CHIVs cannot be neatly placed into any one of the established groups of ssDNA viruses. Furthermore, evidence that RC-Rep genes can be exchanged between unrelated viruses blurs the borders between the major groups of eukaryotic ssDNA viruses and renders the RC-Rep-based classification of the uncultured ssDNA viruses into the circo-, nano- or gemini-like groups obsolete. Indeed, CHIVs with circovirus-like RC-Reps are as similar to circoviruses (i.e., circovirus-like) (32) as they are to tombusviruses (16). Recognizing the limits of the RC-Rep-based approach in classifying uncultured ssDNA viruses, Rosario and colleagues (20) have recently proposed an alternative classification scheme based on a combination of various genomic properties of these viruses. According to the new scheme, viruses are categorised into eight groups (I–VIII) based on their genome orientation, the location of the intergenic region containing the potential stem-loop structure as well as the orientation of the nonanucleotide motif with respect to the RC-Rep gene (20). The diversity of genome organizations observed in CHIVs spans six of the eight proposed groups (Fig. S1 and Table 1), suggesting that such classification scheme might not prove to be practical.

More generally, none of the viral genes taken separately can adequately represent viral history (51), and in the case of viruses with small genomes the whole-genome approaches are also unlikely to be of much value. Indeed, eukaryotic ssDNA viruses, in our opinion, represent a textbook case of organisms for which ancient evolutionary history cannot be reconstructed.

Yet, we consider genes for capsid proteins as more faithful reflections of the latter process (52), and thus advocate a capsid-based classification of these viruses.

Indeed, capsid proteins are hallmarks of viruses and are less likely to leave the realm of virosphere than genome replication proteins that are often exchanged between unrelated viruses, plasmids and cellular chromosomes (53). Consideration of CP genes as markers would also more accurately represent the global diversity of ssDNA viruses—which is likely to be greatly underestimated—since CP genes are more divergent than the corresponding RC-Rep genes (Fig. 4A). According to such capsidocentric virus classification approach (53, 54), CHIVs would be grouped with RNA viruses that possess tombusvirus-like CPs. Intriguingly, such group would traverse two classes of the traditional virus classification scheme proposed by David Baltimore (known as Baltimore classification), in which viruses are grouped into seven classes according to their genome type and the mode of genome replication (55). Unorthodox as it may seem, the above case is not the only example of related viruses having different genome types. For instance, closely related pleomorphic archaeal viruses possess either ss or dsDNA genomes (6), reverse transcribing viruses package either RNA (retroviruses) or DNA (hepadnaviruses) genomes (56), and finally, certain dsRNA viruses have apparently evolved from viruses with ssRNA genomes on several independent occasions (25). Thus, the more we sample the virosphere the more unexpected connections we uncover between viruses that once were considered unrelated.

METHODS

Detection of chimeric viruses in assembled viromes

A set of 103 published viromes available in public databases were downloaded and used in this study. These viromes were obtained from viral communities associated with different types of aquatic samples (freshwater, seawater, hypersaline ponds), eukaryote-associated flora (human gut, saliva, lung, coral, fish) as well as with more peculiar biomes like microbialites or atmospheric samples (Table S1). All viromes were assembled with Newbler 2.6 (454 Life Sciences), with the following parameters: 98% similarity over 35 bp. A BLASTx search was computed to detect contigs containing genes similar to those of RNA viruses (extracted from the NCBI protein database on Aug 2012). Genes were predicted with MetaGeneAnnotator (57) for all contigs that were found to encode putative RNA virus capsid-like proteins (threshold of 50 on bitscore and 0.001 on e-value). Contigs containing at least two genes, one similar to an RNA virus capsid gene and one to an RC-Rep gene were considered as chimeric viruses (Table S1).

Screening of whole-genome shotgun libraries

Different databases from the NCBI were screened for the presence of chimeric viruses based on the 10 capsid proteins from chimeric viruses (the 9 contigs assembled in this study and the BSL_RDHV genome (32)). Searches against Genomic Survey Sequence (GSS), Whole Genome Shotgun (WGS) and High-Throughput Genomic Sequence (HTGS) libraries were performed using tBLASTn, while BLASTp was used to compare CHIV capsid protein sequences to Metagenomic Proteins (env_nr). Putative CHIVs were detected in metagenomes targeting the small eukaryotic fraction in coastal upwelling waters off central Chile (NCBI GI 372349332 and 393314887 ; (38)). Reads from these two datasets were assembled with the same pipeline as the viromes, and three putative CHIV genomes were obtained. Additionally, putative CHIV genome was retrieved from a WGS project of a foraminifera, *Astrammina rara* (NCBI Bioproject PRJNA47149; Contig ADNL01003178; (39)). Chimeric virus sequences were analyzed as described in SI Materials and Methods.

ACKNOWLEDGEMENTS

This work was supported by the French national program Ecosphère continentale et côtière (EC2CO), project CAVIAR (CommunAutés de Virus à ARN). SR was supported by a PhD grant from the French defence procurement agency (DGA, Direction Générale de l'Armement). DV was supported by PHYTOMETAGENE (JST-CNRS), METAPICO (Genoscope), and Micro B3 (funded by the European Union, contract 287589).

REFERENCES

1. King AMQ, Adams MJ, Carstens EB, & Lefkowitz EJ (2011) *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses* (Elsevier Academic Press, San Diego).
2. Mochizuki T, *et al.* (2012) Archaeal virus with exceptional virion architecture and the largest single-stranded DNA genome. *Proc Natl Acad Sci U S A* 109(33):13386-13391.
3. Pietilä MK, Roine E, Paulin L, Kalkkinen N, & Bamford DH (2009) An ssDNA virus infecting archaea: a new lineage of viruses with a membrane envelope. *Mol Microbiol* 72(2):307-319.
4. Tomaru Y, *et al.* (2011) Isolation and characterization of a single-stranded DNA virus infecting *Chaetoceros lorenzianus* Grunow. *Appl Environ Microbiol* 77(15):5285-5293.
5. Yu X, *et al.* (2010) A geminivirus-related DNA mycovirus that confers hypovirulence to a plant pathogenic fungus. *Proc Natl Acad Sci U S A* 107(18):8387-8392.
6. Senčilo A, Paulin L, Kellner S, Helm M, & Roine E (2012) Related haloarchaeal pleomorphic viruses contain different genome types. *Nucleic Acids Res* 40(12):5523-5534.
7. Yamada T, *et al.* (2007) New bacteriophages that infect the phytopathogen *Ralstonia solanacearum*. *Microbiology* 153(Pt 8):2630-2639.
8. Dunlap DS, *et al.* (2013) Molecular and microscopic evidence of viruses in marine copepods. *Proc Natl Acad Sci U S A* 110(4):1375-1380.
9. Holmfeldt K, Odic D, Sullivan MB, Middelboe M, & Riemann L (2012) Cultivated single-stranded DNA phages that infect marine *Bacteroidetes* prove difficult to detect with DNA-binding stains. *Appl Environ Microbiol* 78(3):892-894.
10. Nagasaki K, *et al.* (2005) Previously unknown virus infects marine diatom. *Appl Environ Microbiol* 71(7):3528-3535.
11. Roux S, Krupovic M, Poulet A, Debroas D, & Enault F (2012) Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS One* 7(7):e40418.
12. Rakonjac J, Bennett NJ, Spagnuolo J, Gagic D, & Russel M (2011) Filamentous bacteriophage: biology, phage display and nanotechnology applications. *Curr Issues Mol Biol* 13(2):51-76.
13. Bennett A, McKenna T, & Agbandje-McKenna M (2008) A comparative analysis of the structural architecture of ssDNA viruses. *Computational and Mathematical Methods in Medicine* 9(3-4):183-196.
14. Khayat R, *et al.* (2011) The 2.3-angstrom structure of porcine circovirus 2. *J Virol* 85(15):7856-7862.
15. Chapman MS & Liljas L (2003) Structural folds of viral proteins. *Adv Protein Chem* 64:125-196.
16. Krupovic M (2012) Recombination between RNA viruses and plasmids might have played a central role in the origin and evolution of small DNA viruses. *BioEssays* 34(10):867-870.
17. Ilyina TV & Koonin EV (1992) Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Res* 20(13):3279-3285.
18. Koonin EV & Ilyina TV (1992) Geminivirus replication proteins are related to prokaryotic plasmid rolling circle DNA replication initiator proteins. *J Gen Virol* 73 (Pt 10):2763-2766.
19. Koonin EV (1993) A common set of conserved motifs in a vast variety of putative nucleic acid-dependent ATPases including MCM proteins involved in the initiation of eukaryotic DNA replication. *Nucleic Acids Res* 21(11):2541-2547.
20. Rosario K, Duffy S, & Breitbart M (2012) A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol* 157(10):1851-1871.
21. Delwart E & Li L (2012) Rapidly expanding genetic diversity and host range of the *Circoviridae* viral family and other Rep

encoding small circular ssDNA genomes. *Virus Res* 164(1-2):114-121.
22. Martin DP*, et al.* (2011) Recombination in eukaryotic single stranded DNA viruses. *Viruses* 3(9):1699-1738.
23. Cherwa JE & Fane BA (2011) *Microviridae*: microviruses and gokushoviruses. *Encyclopedia of life sciences*, Chichester, United Kingdom).
24. Krupovic M, Ravantti JJ, & Bamford DH (2009) Geminiviruses: a tale of a plasmid becoming a virus. *BMC Evol Biol* 9:112.
25. Dolja VV & Koonin EV (2011) Common origins and host-dependent diversity of plant and animal viromes. *Curr Opin Virol* 1(5):322-331.
26. Rosario K, Duffy S, & Breitbart M (2009) Diverse circovirus-like genome architectures revealed by environmental metagenomics. *J Gen Virol* 90(Pt 10):2418-2424.
27. Rosario K*, et al.* (2012) Diverse circular ssDNA viruses discovered in dragonflies (Odonata: Epiprocta). *J Gen Virol* 93(Pt 12):2668-2681.
28. Roux S*, et al.* (2012) Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS One* 7(3):e33641.
29. Duffy S, Shackelton LA, & Holmes EC (2008) Rates of evolutionary change in viruses: patterns and determinants. *Nat Rev Genet* 9(4):267-276.
30. Firth C, Charleston MA, Duffy S, Shapiro B, & Holmes EC (2009) Insights into the evolutionary history of an emerging livestock pathogen: porcine circovirus 2. *J Virol* 83(24):12813-12821.
31. Streck AF*, et al.* (2011) High rate of viral evolution in the capsid protein of porcine parvovirus. *J Gen Virol* 92(Pt 11):2628-2636.
32. Diemer GS & Stedman KM (2012) A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol Direct* 7:13.
33. Whon TW*, et al.* (2012) Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J Virol* 86(15):8221-8231.
34. Angly FE*, et al.* (2006) The marine viromes of four oceanic regions. *PLoS Biol* 4(11):e368.
35. Rosario K, Nilsson C, Lim YW, Ruan Y, & Breitbart M (2009) Metagenomic analysis of viruses in reclaimed water. *Environ Microbiol* 11(11):2806-2820.
36. Krupovic M & Forterre P (2011) *Microviridae* goes temperate: microvirus-related proviruses reside in the genomes of *Bacteroidetes*. *PLoS One* 6(5):e19893.
37. Liu H*, et al.* (2011) Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evol Biol* 11:276.
38. Vaulot D*, et al.* (2012) Metagenomes of the picoalga *Bathycoccus* from the Chile coastal upwelling. *PLoS One* 7(6):e39648.
39. Habura A, Hou Y, Reilly AA, & Bowser SS (2011) High-throughput sequencing of *Astrammina rara*: sampling the giant genome of a giant foraminiferan protist. *BMC Genomics* 12:169.
40. Pawlowski J, Holzmann M, Fahrni JF, & Hallock P (2001) Molecular identification of algal endosymbionts in large miliolid foraminifera: 1. Chlorophytes. *J Eukaryot Microbiol* 48(3):362-367.
41. Rochon D, Lommel S, Martelli GP, Rubino L, & Russo M (2011) *Tombusviridae*. *Virus Taxonomy: Ninth Report of the International Committee on Taxonomy of Viruses*, eds King AMQ, Adams MJ, Carstens EB, & Lefkowitz EJ (Elsevier Academic Press, San Diego), pp 1111-1138.
42. Koenig R, Verhoeven JT, Fribourg CE, Pfeilstetter E, & Lesemann DE (2004) Evaluation of various species demarcation criteria in attempts to classify ten new tombusvirus isolates. *Arch Virol* 149(9):1733-1744.
43. Yokoi T, Yamashita S, & Hibi T (2003) The nucleotide sequence and genome organization of *Sclerophthora macrospora* virus A. *Virology* 311(2):394-399.
44. Heller-Dohmen M, Gopfert JC, Pfannstiel J, & Spring O (2011) The nucleotide sequence and genome organization of *Plasmopara halstedii* virus. *Virol J* 8:123.
45. Sherman MB*, et al.* (2006) Removal of divalent cations induces structural transitions in red clover necrotic mosaic virus, revealing a potential mechanism for RNA release. *J Virol* 80(21):10395-10406.
46. Djikeng A, Kuzmickas R, Anderson NG, & Spiro DJ (2009) Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One* 4(9):e7264.
47. Baudoux AC & Brussaard CP (2005) Characterization of different viruses infecting the marine harmful algal bloom species *Phaeocystis globosa*. *Virology* 341(1):80-90.
48. Desnues C*, et al.* (2012) Provirophages and transpovirons as the diverse mobilome of giant viruses. *Proc Natl Acad Sci U S A* 109(44):18078-18083.
49. Park MG, Yih W, & Coats DW (2004) Parasites and phytoplankton, with special emphasis on dinoflagellate infections. *J Eukaryot Microbiol* 51(2):145-155.
50. Speir JA & Johnson JE (2012) Nucleic acid packaging in viruses. *Curr Opin Struct Biol* 22(1):65-71.
51. Koonin EV, Wolf YI, Nagasaki K, & Dolja VV (2009) The complexity of the virus world. *Nat Rev Microbiol* 7(3):250.
52. Krupovic M & Bamford DH (2009) Does the evolution of viral polymerases reflect the origin and evolution of viruses? *Nat Rev Microbiol* 7(3):250.
53. Krupovic M & Bamford DH (2010) Order to the viral universe. *J Virol* 84(24):12476-12479.
54. Abrescia NG, Bamford DH, Grimes JM, & Stuart DI (2012) Structure unifies the viral universe. *Annu Rev Biochem* 81:795-822.
55. Baltimore D (1971) Expression of animal virus genomes. *Bacteriol Rev* 35(3):235-241.
56. Zlotnick A*, et al.* (1998) Shared motifs of the capsid proteins of hepadnaviruses and retroviruses suggest a common evolutionary origin. *FEBS Lett* 431(3):301-304.
57. Noguchi H, Taniguchi T, & Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. *DNA Res* 15(6):387-396.

Vers une disparition des frontières entre groupes ?

D'après l'analyse des 13 nouveaux génomes chimères assemblés, et en particulier l'origine commune des gènes de capside de ces virus, de tels événements de recombinaison entre virus à ARN et ADN apparaissent comme vraisemblablement très rares. A l'inverse, les gènes associés à la réplication (RC-Rep) observés au sein de ces différents génomes ont montré une grande diversité, laissant penser que des événements multiples de recombinaison entre ce gène de capside issu des *Tombusviridae* et différents génomes à ADN simple brin ont eu lieu. L'origine de ce phénomène est sans doute lié à l'adéquation entre ce gène de capside nouveau au sein des petits virus à ADN simple brin, et les multiples versions du gène associé à la réplication : suite au premier événement de transfert, l'association entre les deux gènes n'était certainement pas optimale, et un nouveau génotype présentant une association différente pouvait ainsi être plus compétitif. Dans cette hypothèse, dès que l'une de ces associations serait à même de se maintenir dans l'environnement, un nouveau génotype serait créé, avant d'arriver à une phase d'équilibre entre ces nouveaux virus, leur(s) hôte(s) et leur(s) concurrent(s) (autres virus, parasites, etc). Il est à noter que ces virus chimères semblent étroitement associés aux milieux aquatiques, et plus particulièrement aux cellules algales. Or, plusieurs études ont mis en évidence l'importance des événements d'endosymbiose et de parasitisme intracellulaire pour ce type d'organismes dans les milieux aquatiques (Yoon *et al.*, 2005; Li *et al.*, 2010). Il est ainsi tentant de lier l'apparition de ces virus chimères avec l'existence de ces structures multicellulaires imbriquées en ce que ces structures constituent, au moins en théorie, un espace clos et délimité au sein duquel des génomes viraux issus de virus infectant différents hôtes et sans lien évolutif pourraient être en contact.

Cette analyse met également en exergue les difficultés éprouvées pour reconstruire l'histoire évolutive de structures génétiques simples comme le sont les virus à ADN simple brin. Si le gène codant pour la protéine de réplication était jusqu'ici utilisé comme marqueur phylogénétique, il paraît difficile aujourd'hui de continuer à se baser uniquement sur ces phylogénies pour construire et définir les différents sous-groupes. Rosario et collaborateurs ont ainsi proposé une classification phylogénomique, méthode généralement utilisée pour définir des groupes taxonomiques au sein d'organismes dont les génomes sont soumis à des transferts horizontaux et non assujettis à la reproduction sexuée, comme les micro-organismes (Rosario *et al.*, 2012). Cependant, dans le cas des petits virus à ADN simple brin, les 14 séquences de virus chimères sont d'ores et déjà dispersées dans la grande majorité des catégories ainsi définies (6 sur 8), démontrant les limites de cette analyse.

Afin de mieux caractériser la diversité de ces petits virus à ADN en milieu lacustre, le projet CAVIAR (programme EC2CO 2013, responsable : François Enault) a été mis en place, et vise à caractériser les communautés virales ADN et ARN le long de la colonne d'eau au sein du lac Pavin. Ces nouveaux viromes bénéficiant de la technologie de séquençage à haut-débit HiSeq (Illumina) pourraient apporter des informations décisives concernant la

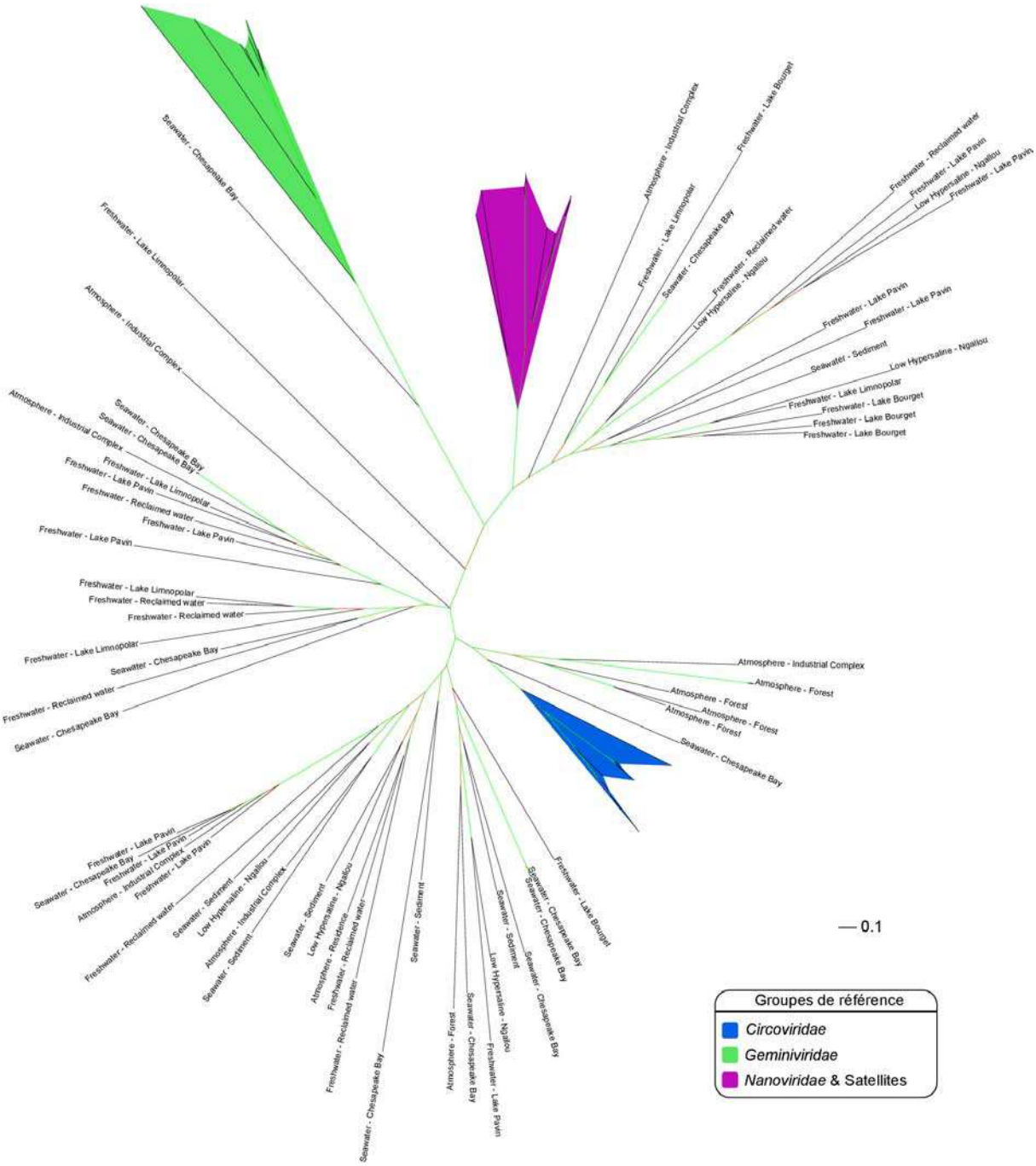


Figure IV.4 : Exemple de phylogénie incluant des séquences métagénomiques basée sur le gène de réplication (RC-Rep). *Les groupes de référence sont surlignés en couleur, et les séquences métagénomiques sont nommées en fonction du jeu de données d'origine. Les branches sont colorées en fonction du score de bootstraps, de rouge (40) à vert (100). Les séquences métagénomiques sont issues des viromes des lacs Pavin et Bourget (Roux et al., 2012), et Limnopolar (Lopez-Bueno et al., 2009), d'eaux usées (Rosario et al., 2009), de la baie de Chesapeake (Rush et al., 2007), et de sédiments marins (Yoshida et al., 2013).*

diversité encore mal caractérisée des virus à ARN dans les milieux lacustre, la diversité des petits virus à ADN simple brin précédemment observée dans les eaux de surface de ce même lac Pavin (Roux *et al.*, 2012), et enfin l'étendue des échanges de gènes entre ces deux fractions virales.

Méta-analyses de viromes pour l'étude des petits virus à ADN simple brin

Les différents résultats présentés ici semblent montrer que les petits virus à ADN simple brin sont plus largement répartis que ne le laissaient penser les observations directes de communautés virales, ce qui est sans doute lié au fait que la taille de ces virions se situe à la limite des observations possibles en microscopie électronique à transmission et en cytométrie en flux. Au-delà de leur diversité, les virus à ADN simple brin semblent également plus complexes qu'initialement imaginé. L'acquisition de gènes par recombinaison semble notamment être relativement fréquente, avec de plus une capacité à s'intégrer au génome de l'hôte (Krupovic *et al.*, 2009; Lefeuvre *et al.*, 2011; Liu *et al.*, 2011). Dans les deux cas étudiés, on peut aussi constater une évolution par accumulation d'insertions dans les parties externes de la protéine de capside. Ce type de phénomène, dont l'intensité est variable en fonction du type d'écosystème étudié plutôt qu'en fonction du sous-groupe analysé, est ainsi vraisemblablement lié aux interactions entre hôtes et virus, plus ou moins fréquentes en fonction des milieux. De manière plus générale, plusieurs études récentes ont mis en évidence une vitesse d'évolution très rapide de ces petits virus à ADN simple brin (Duffy *et al.*, 2008), ce qui, ajouté au potentiel d'acquisition de gènes par recombinaison, fait de ces génomes un terrain propice à l'apparition d'innovations évolutives.

Au-delà des deux groupes analysés dans ce manuscrit (*Microviridae* et virus chimères), les petits virus à ADN simple brin les plus fréquemment retrouvés dans les analyses de viromes peuvent être regroupés sous l'appellation "SCREVs" (Small Circular Rep-Encoding Viruses) dont feraient partie les virus chimères. De manière générale, les génomes de ces virus sont circulaires, et codent pour deux gènes principaux : l'un associé à la réplication, l'autre à la formation de la capside. De telles entités virales simplifiées à l'extrême étaient, jusqu'à l'apparition des approches métagénomiques, classées au sein de familles (*Circoviridae*, *Geminiviridae*, et *Nanoviridae*) relativement bien définies sur la base d'analyses phylogénétiques du gène associé à la réplication. Toutefois, les études récentes ont mis en évidence une diversité exceptionnelle de ces structures génomiques (Rosario *et al.*, 2009b; Kim *et al.*, 2011; Delwart & Li, 2012; Ng *et al.*, 2012). Ces nouveaux génomes ne

correspondent, pour la plupart, à aucun groupe connu, et leur diversité est tellement importante qu'il est impossible à l'heure actuelle de détecter un nombre raisonnable de sous-groupes cohérents au sein d'une phylogénie incorporant ces séquences environnementales (Figure IV.4).

Pour dépasser cette limite, une classification phylogénomique a été proposée par Rosario et collaborateurs. Mais comme le montre l'exemple des virus chimères, ces petits virus à ADN simple brin semblent présenter, en plus d'une vitesse d'évolution rapide (proche de celle des virus à ARN simple brin), une forte capacité de recombinaison et donc de transfert de gène horizontal. En considérant que ces virus peuvent de plus être intégrés au sein des génomes, et potentiellement échanger des gènes avec des structures de type plasmides ou transposons (Krupovic *et al.*, 2009), l'idée de reconstituer leur histoire évolutive sous la forme d'un arbre binaire paraît utopique. Au final, il est possible que la reconstruction de l'histoire évolutive de tous ces petits virus soit hors de notre portée de par le fait que la quantité d'information contenue dans le génome de ces virus est simplement trop limitée, et qu'il faille se contenter de définir des groupes de virus similaires par la comparaison de génomes complets, et reconstituer uniquement leurs relations évolutives récentes. Alternativement, il est possible d'étudier chaque gène indépendamment, en considérant que leur association au sein des génomes est trop faiblement contrainte pour générer une véritable évolution cohérente. Dans ce cas, il serait possible de reconstituer la trajectoire de chaque gène au sein de différents éléments de réplications, et donc son histoire évolutive propre.

L'analyse transversale d'un ensemble de viromes publiés peut ainsi révéler de nouvelles informations importantes sur des groupes viraux particuliers, concernant à la fois leur diversité, leur distribution dans les différents environnements, leur(s) hôte(s) potentiel(s), ou encore les mécanismes d'évolution opérant au sein de leurs génomes. Cette approche de méta-analyse de viromes permet ainsi de compléter les informations obtenues sur ces groupes viraux d'intérêt *via* les approches d'isolement et culture ou d'amplification de gènes marqueurs. L'un des atouts principaux de ces analyses de viromes est la disponibilité de génomes complets, plus informatifs quant il s'agit d'étudier la diversité virale notamment pour déterminer l'existence de nouveaux groupes ou d'éventuels transferts de gènes horizontaux.

Conclusion

Les travaux réalisés au cours de cette thèse sur l'étude des communautés virales par approche métagénomique ont effectivement permis de mieux caractériser la distribution des virus, leur diversité génétique et génomique, et certains de leurs mécanismes d'évolution.

Dans un premier temps, le développement d'outils bioinformatiques a permis de mettre en place le serveur Metavir, qui constitue aujourd'hui l'un des outils de référence dans le domaine du traitement des données de métagénomique virale (Chapitre I, (Fancello *et al.*, 2012a; Reyes *et al.*, 2012)). Metavir est à l'heure actuelle le seul serveur web proposant l'analyse complète d'un virome (hors assemblage), quel que soit le type de séquençage utilisé et la taille des séquences étudiées. Plus de 2 000 viromes ont été analysés depuis sa mise en service en septembre 2011, que ce soit dans le cadre d'approches écologiques (Yoshida *et al.*, 2013) ou épidémiologiques (Pérez-Brocal *et al.*, 2013). De plus, un certain nombre de développements sont envisagés, comme une meilleure définition des gènes marqueurs proposés de manière à mieux couvrir l'ensemble des génomes viraux connus, ou encore l'association d'indices numériques de diversité aux courbes de raréfaction, dans le but d'offrir le service le plus complet possible.

Les différentes analyses de viromes conduites ont globalement permis de mieux caractériser la diversité taxonomique et fonctionnelle des communautés virales aquatiques autour du globe. L'analyse et la comparaison de ces génomes viraux de l'environnement a notamment révélé la présence de plusieurs gènes impliqués dans différents cycles métaboliques cellulaires (Chapitre II), ou encore confirmé la capacité de dispersion globale des particules encapsidées (Chapitre III). Dans ce cadre, un certain nombre de paramètres, et en premier lieu la salinité du milieu, ont pu être mis en exergue en tant que facteurs structurant ces communautés virales (Chapitre III).

Enfin, ces approches de métagénomique virale ont également permis de lever le voile sur des pans entiers de la virosphère peu considérés jusqu'à maintenant, comme les petits virus à ADN simple brin (Chapitre IV). Ces derniers semblent en effet bien plus diversifiés et complexes en terme de cycle d'infection et mécanismes évolutifs que ne le laissaient supposer les souches isolées, et pourraient se révéler potentiellement importants pour des questions de santé ou au sein des écosystèmes.

L'ensemble de ces travaux et résultats ont ainsi apporté différents éléments pour appréhender des questions fondamentales plus larges, à l'heure actuelle encore en débat, concernant l'histoire évolutive et la nature des virus d'une part, et leur place au sein des écosystèmes d'autre part.

Nature des virus et mécanismes d'évolution au sein de la virosphère

Les virus occupent une place à part dans l'histoire du vivant, et les questions de leur origine et de leur rôle au cours de l'évolution de la vie sur Terre sont aujourd'hui encore en débat. Cette spécificité des virus relève de plusieurs facteurs, et notamment du fait que les virus, de par leur structure simple, défient par nature les conceptions établies au sujet de la vie sous forme cellulaire (Forterre, 2006).

La place du monde viral par rapport au monde cellulaire

Les virus présentent tous un style de vie dit “parasite”, et ne sont donc pas autonomes comme peuvent l'être la plupart des être vivants cellulaires. Certains des éléments leur manquant sont parmi les plus fondamentaux dans le cadre de la définition de la vie, notamment les composants de la machinerie permettant la synthèse protéique (ARN ribosomal et protéines associées). Ces observations ont conduit certains auteurs à écarter les virus du monde du vivant (Van Regenmortel, 2003; Moreira & López-García, 2009). Dans cette optique, les virus sont considérés comme des vecteurs d'information génétique, inertes en l'absence de la machinerie cellulaire de l'hôte. En ce sens, ils se rapprochent d'autres éléments génétiques mobiles comme les transposons ou les plasmides, avec qui les virus partagent certaines caractéristiques.

Rejetant l'argument du style de vie parasitaire des virus comme les excluant *de facto* du règne du vivant, certains auteurs définissent un état de “virocell”, ou “usine à virions”, qui correspond au moment où les capacités de la cellule sont entièrement détournées par et pour le virus (Forterre, 2013). Considérant la “virocell” comme la forme vivante du virus, ces derniers sont alors inclus au sein du vivant.

Enfin, certains auteurs classent les virus dans une zone grise entre le vivant et le non-vivant, comme “assemblage doué de potentiel de vie”, faisant l'analogie avec la graine d'un organisme végétal, pas tout à fait vivante, mais pouvant le devenir si placée dans de bonnes conditions (Villarreal, 2004).

Les différents résultats d'analyse de viromes ont apporté des informations supplémentaires à ce débat (Chapitre II, (Sharon *et al.*, 2011; Thompson *et al.*, 2011)). La détection de plus en plus fréquente de gènes associés au métabolisme cellulaire au sein de certains virus et au cours des analyses de viromes, et le fait que cette présence de gènes métaboliques semble être un phénomène général (au moins au sein des virus à ADN), semble ainsi contradictoire avec la vision “inerte” du monde viral. L'origine de ces gènes similaires

entre génomes cellulaires et viraux est encore en débat. Si l'on considère les génomes viraux comme parasites des génomes cellulaires, on peut alors imaginer que ces gènes sont d'origine cellulaires, et "volés" par le virus dans son propre intérêt. A l'inverse, si ces gènes sont d'origine virale, les virus deviendraient alors de véritables créateurs de diversité génétique et d'innovations évolutives.

Les analyses de viromes indiquent également que certains génomes viraux codent pour des protéines ribosomales (Chapitre II, (Sharon *et al.*, 2011)). Ainsi, et malgré la caractérisation encore largement incomplète du monde viral, seule la présence d'ARN ribosomal semble à l'heure actuelle pouvoir distinguer les génomes cellulaires des génomes viraux. Cette absence de description de fragments génomiques identifiés comme d'origine virale et contenant un gène codant pour l'un des ARN ribosomaux, malgré l'application des approches métagénomiques et l'utilisation des techniques de séquençage à haut-débit, laisse à penser que les génomes viraux comprenant des ARN ribosomaux sont soit extrêmement rares, soit inexistants, et que les ARNr constituent les marqueurs principaux de différenciation entre les organismes cellulaires et les virus. Ainsi, en considérant une définition du vivant axée sur l'existence d'une unité de réplication à support nucléotidique se transmettant dans le temps et l'espace au sein de la biosphère, les virus constituent bien un monde à part, un monde encapsidé évoluant en parallèle et en interaction avec le monde cellulaire (Raoult & Forterre, 2008). La frontière de la vie serait alors déplacée au niveau des prions, exclus du vivant de par l'absence de matériel génétique, et aux GTA, pour lesquels il n'existe pas d'unité de réplication identifiée.

Origine évolutive des virus

Une autre question fondamentale encore en débat à l'heure actuelle est celle de l'origine des virus, même si plusieurs éléments sont aujourd'hui largement acceptés. Tout d'abord, la plupart des auteurs considèrent aujourd'hui que les virus ont probablement une origine extrêmement ancienne et, au moins pour certains d'entre eux, que cette origine se trouve au niveau de l'ancêtre commun des lignées cellulaires actuelles. Dans ce cadre, l'analyse des *Microviridae* à partir des séquences de viromes a ainsi montré que ces derniers étaient certainement apparus avant la différenciation des différentes lignées de bactéries, et possédaient une origine bien plus ancienne qu'initialement pensée (Chapitre IV). Ensuite, l'origine des virus est maintenant reconnue comme polyphylétique, et il n'existe ainsi probablement pas d'ancêtre commun à l'ensemble du monde viral, ce qui justifie à la fois

l'existence d'hypothèses multiples sur l'origine des différentes familles virales et l'hétérogénéité de ces dernières.

Une première hypothèse est celle de l'évolution “progressive”, dans laquelle les virus seraient des éléments génétiques mobiles ayant acquis la possibilité de se transmettre de cellule en cellule *via* une capside. Les *Geminiviridae* par exemple constituent l'une des familles virales susceptibles d'être apparue à partir d'un élément mobile (Krupovic *et al.*, 2009). Les similarités dans le fonctionnement moléculaire et dans certains cas au niveau des séquences entre certains rétrovirus et des rétro-éléments mobiles de génomes eucaryotes semblent plus généralement attester d'un lien évolutif entre ces deux types d'entités génétiques. L'exemple des virus chimères ADN-ARN semble correspondre à cette hypothèse d'évolution progressive et de recombinaison entre une unité de réplication et un gène de capside (Chapitre IV, (Diemer & Stedman, 2012)). De même, des liens évolutifs importants avec les éléments mobiles comme les plasmides ont été effectivement observés lors de l'analyse de viromes hypersalins du programme Archevir (Chapitre III). Ainsi, s'il n'est pas possible de généraliser ce modèle à l'ensemble du monde viral, il semble qu'une partie au moins des virus puissent avoir une origine liée aux éléments génétiques mobiles.

L'hypothèse inverse place l'origine des virus dans le monde cellulaire, avec un mouvement de réduction et de simplification du génome qui aurait conduit cette cellule du monde membranaire vers le monde encapsidé (*via* un processus de dégénérescence cellulaire). L'existence de virus géants comme le groupe des NCLDV *(nucleocytoplasmic large DNA virus,* dont font partie les *Poxvirus* et *Mimivirus*), dont la complexité est quasiment équivalente à celle d'une petite bactérie, semble être un élément en faveur de cette théorie. Ainsi, un organisme cellulaire pourrait être passé par un stade de symbiose ou de parasite intracellulaire obligatoire, avant de devenir un virus à part entière par l'acquisition de la capacité à former une capside (Claverie & Abergel, 2013). Un résultat proche voire similaire pourrait toutefois être atteint *via* un mécanisme légèrement différent, au sein duquel un virus pourrait être créé à partir d'éléments du génome cellulaire s'associant pour créer une entité indépendante (théorie dite de l'échappement cellulaire).

Si ces deux théories se basent sur la pré-existence d'entités cellulaires avant l'apparition des virus, certains auteurs proposent au contraire une antériorité des structures virales sur les cellules (Koonin *et al.*, 2006). Les virus sont alors vus comme des descendants directs d'éléments génétiques primaires existant au sein d'un monde pré-cellulaire. L'apparition des capsides virales aurait précédé l'apparition des cellules “modernes” (à membrane lipidique), même si ces éléments encapsidés ne peuvent pas être qualifiés à proprement parler de “virus” puisqu'ils ne pouvaient infecter de cellules. Cette hypothèse est soutenue par l'existence de gènes conservés spécifiques des génomes viraux, observation

également retrouvée lors de l'analyse des viromes (Chapitre III). Ces virus ancestraux pourraient alors être considérés comme constituant un quatrième domaine au sein des êtres vivants (aux côtés des eucaryotes, bactéries, et archées). Sans formuler le postulat d'une antériorité des structures virales, Nasir et collaborateurs ont récemment proposé l'existence d'un super-groupe de virus constitué des virus géants à ADN double brin ayant coexisté avec l'ancêtre commun des organismes cellulaires (LUCA), qui formeraient ainsi un groupe parallèle aux trois domaines de la vie cellulaire (Nasir *et al.*, 2012).

Ainsi, si la distinction entre un monde encapsidé et un monde à ribosome semble de plus en plus claire, la ou les origines des différents groupes viraux restent encore à caractériser.

Enfin, quelle que soit la nature (vivante ou non) et l'origine des virus, la plupart des auteurs estiment qu'ils ont joué un rôle important voire décisif dans l'apparition de plusieurs éléments fondamentaux de la vie cellulaire (Brüssow, 2009). Ainsi, les premiers éléments de nature ADN au sein du monde ARN pourraient être apparus chez des virus infectant des cellules à génome ARN (Forterre, 2006). Cette origine virale des machineries génomiques à ADN pourrait ainsi être à l'origine des différences observées entre les systèmes des trois grandes lignées cellulaires modernes (bactéries, archées et eucaryotes). D'autres théories sur l'évolution fondamentale des cellules ont impliqué les virus, comme l'hypothèse d'une origine virale pour le noyau, dans laquelle un virus à ADN double brin (de type Poxvirus) aurait été intégré à une cellule par symbiose (Takemura, 2001; Claverie, 2006; Forterre, 2006; Bell, 2009). Les virus pourraient également être impliqués dans la formation d'autres organites, tels que les mitochondries ou les chloroplastes, et plus globalement dans la plupart des innovations évolutives (Claverie, 2006; Forterre & Prangishvili, 2009; Villarreal & Witzany, 2010).

Définition et classification des virus

La caractérisation de plus en plus avancée de la diversité génomique virale, notamment *via* les approches métagénomiques, et la description de structures à la fois de plus en plus complexes (virus géants) et simples (SCREVs) a considérablement réduit les caractéristiques spécifiques du monde viral. En l'état actuel des connaissances, on peut considérer que (i) la constitution d'une unité de réplication (le génome viral), (ii) la présence sur ce génome d'un gène de capside (et la possibilité associée de générer sa propre capside), et (iii) l'absence de ribosome, forment les éléments de base communs à l'ensemble du monde

viral. Ces éléments constitutifs des virus permettent d'exclure du monde viral les GTA, qui malgré leur capside sont exclus de par l'absence d'unité de réplication formelle, et les éléments génétiques mobiles autonomes comme les plasmides et transposons qui à l'inverse sont exclus de par l'absence de capside. Enfin, certaines structures aux frontières du monde viral sont également écartées, comme les viroïdes, formés d'ARN nu et sans capside, et les prions qui ne contiennent pas d'acide nucléique. A l'inverse, les virophages et virus satellites sont bien inclus dans cette définition du virus, même si leur cycle de développement semble associé non pas à un hôte cellulaire, mais à un autre virus, et leur classification encore contestée (Krupovic & Cvirkaite-Krupovic, 2011; Desnues & Raoult, 2012; Fischer, 2012).

La reconstitution de l'histoire évolutive des virus a ainsi été ébauchée à partir de différents gènes marqueurs, notamment les gènes codant pour les fonctions essentielles comme la réplication ou l'encapsidation du génome. Si ces approches ont permis d'établir l'histoire évolutive de familles virales particulières, comme par exemple pour les *Microviridae,* le seul élément commun à l'ensemble des virus et qui pourrait donc constituer un élément de classification virale universelle reste la capside. Dans ce cadre, s'il n'existe pas de signal au niveau des séquences nucléotidiques ou protéiques, un certain nombre de lignées (non nécessairement liées entre elles) ont été définies sur la base des différentes conformations structurales des protéines de capsides au sein des virions (Krupovic & Bamford, 2009; Abrescia *et al.*, 2012), et il devrait être possible de reconstituer l'histoire de chacune d'entre elles. Toutefois, l'exemple des virus chimères (Chapitre IV, (Diemer & Stedman, 2012)), au sein desquels la protéine de capside est transférée d'un génome à l'autre, ou encore la détection de la protéine majeure de capside du *Salterprovirus His1* dans différents génomes (Chapitre III), rappellent qu'il n'existe aucun marqueur “parfait” au sein du monde viral.

La classification des groupes viraux et la reconstitution de leur histoire évolutive pourrait alors passer par des analyses phylogénomiques plutôt que par l'analyse de séquences de gènes uniques. De telles approches ont été réalisées avec succès sur des cyanophages (Ignacio-Espinoza & Sullivan, 2012) ou sur des virus eucaryotes (de Villiers *et al.*, 2010), mais pourraient s'avérer plus délicates voire impossibles à appliquer pour les structures génomiques plus simples comme les petits virus à ADN ou ARN, ainsi que semble le montrer l'exemple des virus chimères (Chapitre IV). En effet, le niveau de diversification au sein de ces groupes et la petite taille des génomes limitant la quantité d'information disponible compliquent la reconstitution de trajectoires évolutives claires. Ainsi, l'établissement d'une histoire évolutive ancienne de l'ensemble des lignées virales est probablement impossible, de par le fait que le signal évolutif observé à travers l'analyse des génomes ou des structures de capside est saturé par les évolutions multiples et rapides que subissent les entités virales.

Les approches de métagénomique virale ont ainsi permis une caractérisation plus complète de la virosphère, et apporté un certain nombre d'éléments nouveaux et inattendus ayant fortement modifié la perception des communautés virales. Cette caractérisation des virus de l'environnement devrait également permettre une meilleure intégration de ces entités au sein des modèles écologiques, de manière à mieux comprendre les impacts multiples des virus au niveau des écosystèmes.

Place et influence des virus dans les écosystèmes

Si les viromes ont permis de mieux décrire la richesse génétique des génomes viraux de l'environnement et ont apporté des connaissances tout à fait nouvelles sur ces communautés, un certain nombre de limites sont néanmoins apparues lors de l'application de ces approches à l'écologie virale. En particulier, les aspects de contamination des jeux de données, l'absence de caractère quantitatif des viromes, ou encore la faible taille des séquences étudiées limitent fortement les conclusions issues de ces analyses. L'absence de lien entre les trois principaux éléments constitutifs des virus que sont leur génome, la structure de leur capside, et leur spectre d'hôtes constitue également un écueil important de l'écologie virale aujourd'hui. En effet, ces trois paramètres sont cruciaux pour l'intégration des virus au sein des modèles écologiques, mais ils sont à l'heure actuelle étudiés séparément, et très compliqués à associer en l'absence d'isolement du virus.

Un certain nombre de méthodologies actuellement en développement devraient toutefois progressivement permettre de dépasser ces différentes limites. Associées au développement des outils de préparation et d'analyse des viromes, ces méthodes devraient ainsi proposer une vision plus intégrée et complète des communautés virales environnementales.

Perspectives et développements des viromes

L'efficacité des protocoles d'extraction des particules virales étant en constante amélioration, ces derniers devraient apporter dans les années à venir la possibilité de véritablement purifier les capsides virales à partir d'échantillons complexes, et ainsi limiter au maximum la contamination des viromes par des génomes cellulaires (Duhaime & Sullivan, 2012; Willner & Hugenholtz, 2013). Ces développements devraient ainsi permettre d'étendre

les approches de métagénomique virale aux systèmes tels que les sols, pour lesquels l'analyse de viromes s'est heurtée aux difficultés liées à l'extraction des capsides virales à partir de matrices solides.

De plus, les données actuelles de métagénomique ne peuvent pas être considérées comme strictement quantitatives, de par les biais multiples introduits notamment par l'étape d'amplification aléatoire. Les analyses de viromes indiquent donc au mieux une tendance générale des différentes quantités de chaque type de virus présents initialement dans l'échantillon sous forme de virion. Or, il est primordial de disposer d'approches quantitatives pour étudier les virus avec une vision écologique. Des développements sont ainsi en cours pour réaliser de véritables viromes quantitatifs, même si ces derniers sont pour l'instant limités à la fraction des virus à ADN double brin (Duhaime *et al.*, 2012). Différentes méthodes d'analyses complémentaires, comme la PCR quantitative ou les puces à ADN, devraient aussi être utilisées, en partie sur la base de résultats d'analyses métagénomiques (Aw & Rose, 2012).

Les évolutions successives des techniques de séquençage laissent également présager d'un rapprochement entre la métagénomique et la génomique pour ce qui est de l'étude des virus. Ainsi, les viromes seront constituées de génomes complets (ou quasi-complet) assemblés, de telle sorte que la métagénomique virale sera alors véritablement une étude de génomique des communautés au sens propre du terme, plutôt que l'analyse d'un pangénome communautaire fragmenté tels qu'étaient les premiers jeux de données. Au-delà de l'étude écologique des communautés virales, la réduction des coûts et l'accélération des procédures de préparation des viromes (notamment l'étape de séquençage) feront certainement des viromes des outils de diagnostic majeurs dans les années à venir, à condition notamment que les outils d'analyse bioinformatique continuent à gagner en précision et en accessibilité (Barzon *et al.*, 2013; Bibby, 2013).

Certains outils des sciences de l'information, initialement développés en dehors du champ de la biologie, pourraient alors se révéler particulièrement utiles pour traiter les masses de données que constitueront les viromes, notamment dans le cadre d'analyses comparatives. En effet, le traitement et la visualisation de telles quantités de données éprouvent aujourd'hui les limites des outils bioinformatiques classiquement utilisés, et des approches de fouille de données, réseau, et apprentissage automatique seront de plus en plus indispensables pour la réalisation d'analyses dans un temps raisonnable (Huttenhower & Hofmann, 2010; Yip *et al.*, 2013). Ces techniques permettront notamment de traiter des masses de données considérables au sein d'analyses comparatives, de manière à isoler certains groupes d'intérêt sur la base des paramètres d'accompagnement des échantillons, ou encore de déduire des règles

d'associations entre les différents éléments étudiés, qu'il s'agisse de différents virus, différentes populations, ou encore des communautés virales et cellulaires.

A l'inverse, certains développements bioinformatiques initialement dédiés aux viromes pourraient s'appliquer plus largement aux études d'écologie microbienne. À titre d'exemple, l'analyse conjointe de différents gènes marqueurs par des approches métagénomiques semble mieux restituer la composition des communautés que les analyses d'amplicons basées sur une approche d'amplification PCR d'un seul gène marqueur (Annexe 1 : (Roux *et al.*, 2011; Klingenberg *et al.*, 2013)). De même, la transition entre la métagénomique actuelle et la génomique des communautés sera effective en premier lieu pour les viromes, de par la taille des génomes viraux, mais interviendra également à terme pour les fractions cellulaires, comme le démontrent les premiers projets à grande échelle de métagénomique bactérienne (Arumugam *et al.*, 2011). Dans ce cadre, les outils initialement développés pour l'écologie virale pourraient être adaptés par la suite pour des études plus générales d'écologie microbienne.

Liens entre séquences métagénomiques, observations directes et hôtes des virus

L'intégration des virus aux études écosystémiques est le plus souvent réalisée sur la base d'observation de communautés complètes, séparées sur des critères de taille de la capside. Les dynamiques des différentes populations observées sont alors corrélées entre elles et avec différents paramètre environnementaux, afin d'en déduire des interactions potentielles (Jacquet *et al.*, 2005; Sime-Ngando *et al.*, 2007; Bettarel *et al.*, 2011). L'une des limites de ces études reste l'absence d'identification formelle des membres de la communauté, qui empêche de véritablement comprendre quels sont les acteurs clés de ces cycles. L'application des approches métagénomique a permis de mieux décrire les communautés virales, notamment du point de vue génétique, mais il n'est malheureusement pas possible actuellement de lier les séquences et les populations observées, de telle sorte que la dynamique des communautés virales n'est pour l'instant étudiée qu'à un niveau très large.

De même, l'observation de formes de capsides exceptionnelles a récemment suscité un regain d'intérêt pour les communautés virales de certains milieux spécifiques, comme les milieux de forte salinité (Sime-Ngando *et al.*, 2010). La réalisation de viromes à partir de ce type d'échantillon devait permettre de caractériser ces virus aux morphologies tout à fait nouvelles. Cependant, ces analyses se sont révélées inefficaces pour l'étude de groupes de virus pour lesquels aucune séquence de référence n'est disponible (Chapitre III). En effet, les analyses naïves comme les approches de modélisation de structures tri-dimensionnelles ou la

détection de domaines fonctionnels peuvent s'appliquer à des séquences relativement lointaines des séquences connues, mais ne peuvent rien pour les séquences et familles de protéines totalement nouvelles.

Différents protocoles sont ainsi en cours de développement afin de réaliser une étude par "compartiment" viral, en couplant les approches métagénomiques à différentes étapes de préparation et séparation des fractions au sein de l'échantillon. De telles approches de métagénomique ciblée peuvent être basées sur une détection du groupe d'intérêt par observation en microscopie, par PCR, ou par séparation des différents génomes viraux en PFGE, et pourraient ainsi permettre d'associer un virome à une sous-population précise et identifiée de virus (Bergeron *et al.*, 2007; Brum *et al.*, 2013). Ce type de sélection peut être menée jusqu'à l'obtention et le séquençage d'un virus unique, ce qui permettrait de disposer plus aisément du génome complet d'un virus d'intérêt (Allen *et al.*, 2011).

Au-delà d'une description plus complète des communautés virales, l'un des points cruciaux pour pouvoir véritablement considérer les virus en terme d'écologie réside en la mise au point de méthodes d'observation *in situ* et d'identification des hôtes (Willner & Hugenholtz, 2013). Il est parfois possible d'utiliser des comparaisons de séquences pour identifier au sein des viromes des séquences proches de motifs de type CRISPR ou de prophages, et ainsi formuler des hypothèses concernant l'hôte du virus étudié (Anderson *et al.*, 2011; Williamson *et al.*, 2012), mais ces approches restent limitées par le nombre de références disponibles et peu robustes en l'état actuel des bases de données.

Certains développements méthodologiques très récents, comme la technique PhageFISH (ou hybridation *in situ* fluorescente) appliquée aux génomes viraux pourraient ainsi être utilisées pour identifier le ou les hôtes des virus marqués par l'association d'un deuxième marquage ciblant les communautés d'hôte potentiels (Allers *et al.*, 2013). Une approche telle que le " Viral Tagging " permet également de sélectionner des virus d'intérêt au sein de la communauté sur la base de leur capacité d'attachement à une cellule d'intérêt (Deng *et al.*, 2012). Couplées aux approches de type *single-cell* et à des observations en microscopie, il est alors possible en théorie de caractériser l'ensemble d'une communauté virale associée à un hôte d'intérêt, tant morphologiquement que génétiquement. Une étape d'isolement et de culture reste toutefois indispensable pour décrypter au mieux les interactions hôtes-virus.

Intégration des virus aux modèles écologiques

L'ensemble des ces informations concernant notamment la quantification des différents virus et leur spectre d'hôte est primordial pour pouvoir considérer au mieux la place

des virus dans les écosystèmes. En effet, au-delà de leur impact au niveau des producteurs primaires et en terme de mortalité de l'hôte, les virus interviennent très certainement à des niveaux multiples et sur l'ensemble des organismes.

L'impact des communautés virales sur les cycles biogéochimiques majeurs dépasse tout d'abord certainement le cas du “viral shunt” et le relargage de nutriments sous forme de matière organique dissoute. Si, comme semblent l'indiquer les différentes analyse de viromes, les virus procèdent à une véritable reprogrammation métabolique de la cellule hôte durant l'infection (Chapitre II), et que cette dernière n'est pas seulement transitoire mais peut s'inscrire durablement dans le temps, ces modifications des cycles cellulaires considérés au niveau des populations entraînera alors certainement une modification importante du fonctionnement de l'écosystème.

Au-delà de la composition des communautés, l'impact des virus est également à prendre en compte pour l'évolution des génomes de leur hôtes. Dans ce cadre, les interactions génétiques entre le monde cellulaire et le monde viral restent également encore à éclaircir. La diversité des systèmes CRISPR-Cas par exemple, est encore loin d'être caractérisée (Anderson *et al.*, 2011). Ainsi, un système “inverse” Cas-CRISPR vient d'être découvert (Seed *et al.*, 2013), et laisse imaginer un panel de mécanismes d’interactions entre phages et micro-organismes procaryotes plus large et complexe que les connaissances actuelles ne le laissent penser. Les différentes interactions pouvant exister entre virus et cellule hôte et leur fréquence d'apparition au sein des populations naturelles sont donc encore certainement pour la plupart à caractériser.

Un aspect supplémentaire de l'intégration des virus dans le fonctionnement des écosystèmes réside dans le lien entre les aspects de transfert de gènes et recombinaison avec les données de plus en plus nombreuses concernant le fort taux de parasitisme et de symbiose existant dans les écosystèmes naturels. Slimani et collaborateurs ont par exemple récemment décrit le “champ de bataille” que peut représenter une cellule d'amibe pour différents micro-organismes bactériens et viraux, y compris des virus géants et des virophages (Slimani *et al.*, 2013). Cet ensemble de cellules imbriquées pourrait être à l'origine de l'apparition des virus chimères, en ayant permis le contact entre un virus à ADN simple brin infectant une algue et un virus à ARN infectant un champignon parasite d'algue (Chapitre IV). L'existence de telles zones regroupant dans un milieu confiné différents types de virus, infectant potentiellement différents hôtes, pourrait donc donner lieu à des échanges de gènes entre différents génomes, viraux ou cellulaires. Ainsi, les phénomènes de parasitisme et de symbiose pourraient être décisifs dans l'histoire évolutive des génomes viraux et cellulaires. Associée à la distribution ubiquiste des virus telle qu'elle semble exister, cette prise en compte de parasitismes multiples

fait des virus un vecteur unique de transfert de gènes à travers l'espace et les grands domaines de la vie.

De manière générale, il existe ainsi une gamme d'action des virus sur leur hôte extrêmement large, d'un état quasiment similaire à un mutualisme aux infections terriblement létales pour les populations d'hôtes. A l'image du spectre de stratégies évolutives décrites pour les organismes supérieurs, il est raisonnable d'imaginer qu'au sein du monde viral, un ensemble continu de cycles de reproduction se soit également développé. Si les interactions associées aux cycles les plus infectieux sont de mieux en mieux caractérisées, il est toutefois fort probable que les communautés virales environnementales recèlent un nombre important de virus moins efficaces en terme de mortalité, mais potentiellement tout aussi important en terme d'influence sur l'évolution, la diversité et la structure des communautés d'organismes cellulaires.

Fonctions potentielles de la "Matière Noire Virale"

Les différentes analyses de viromes environnementaux ont systématiquement pointé la grande richesse génétique des communautés virales, tout en constatant que la plupart des séquences ne correspondaient à aucune référence connue (Chapitre III, (Edwards & Rohwer, 2005)). Plus généralement, pour les 142 viromes publics disponibles sur le serveur Metavir, seulement 10.5 % (± 10 %) des séquences présentent une similarité significative avec un génome viral séquencé. Si cet état de fait est en partie lié à la taille limitée des séquences métagénomiques (Chapitre III, (Wommack *et al.*, 2008)), l'analyse de viromes issus des technologies de séquençage les plus récentes a montré que la plupart des gènes prédits, bien que visiblement complets, étaient effectivement absents des génomes viraux actuellement connus (Chapitre III, (Emerson *et al.*, 2012; Minot *et al.*, 2012b)). Ainsi, parmi les différents projets publics formés de séquences assemblées sur le serveur Metavir, la proportion de gènes affiliés varie entre 16 et 25 % en fonction du type de milieu étudié.

Cette richesse en gènes apparemment sans fin des communautés virales environnementales a mené au développement du concept de "Matière Noire Virale" pour désigner cet ensemble exceptionnel de gènes nouveaux et non caractérisés. Ces gènes sont généralement considérés comme tous fonctionnels et importants pour les virus. En effet, il est généralement admis que les organismes parasites vont subir une réduction de leur génome, avec une sélection des petits génomes de par leur plus grande efficacité de réplication en terme de coût énergétique, même si ce dernier point reste encore en débat (Moran, 2002). De même, il est généralement considéré qu'une forte pression de sélection s'exerce sur la taille

des génomes viraux, de sorte que les parties non fonctionnelles seraient réduites au minimum. Selon cette hypothèse, l'ensemble de la Matière Noire Virale (ensemble des gènes prédits au sein des génomes et métagénomes viraux dont la fonction est inconnue) constituerait donc bien un ensemble de gènes fonctionnels incroyablement diversifiés (Hurwitz & Sullivan, 2013).

Pourtant, certains éléments semblent aller à l'encontre de cette vision du génome viral comme entièrement codant et transcrit. En effet, la taille d'un génome viral semble avant tout contraint par la conformation de la capside, et donc indirectement par les gènes responsables de cette encapsidation. Dans ce contexte, le cas des virus chimères, pour lesquels un transfert de la protéine codant pour la capside est observé, est particulièrement illustratif (Chapitre IV). Si les génomes de ces virus chimères semblent posséder le même contenu génomique que les virus à ADN simple brin avec lesquels il partagent le gène associé à la réplication ainsi que la nature de génome, leur taille est plus importante, et plutôt proche de celle des *Tombusviridae*, dont ils partagent la capside.

Cette observation peut être associée aux gènes additionnels observés sur ces génomes chimères : aucun de ces gènes n'a pu être affilié à une séquence connue, et de plus aucune similarité n'a pu être observée entre ces différents virus pour ces mêmes gènes. Ces gènes additionnels prédits peuvent effectivement coder pour des protéines encore inconnues, et constituer une partie variable au sein de ces génomes. Toutefois, il est également possible d'imaginer que cette partie du génome n'est tout simplement pas fonctionnelle, mais que le gain en terme de survie et de multiplication (*i.e.* le gain de fitness) associé à la réduction de la taille du génome n'est pas assez important pour que les mutations menant à cette réduction soient fixées dans la population de virus chimères. Cet exemple est particulièrement intéressant puisque la miniaturisation du génome et son extrême efficacité de codage sont généralement associées à ce type de petits virus. Il se pourrait ainsi que de telles trajectoires évolutives (réduction à l'extrême du génome par l'intermédiaire d'une réduction de l'espace intergénique, de gènes chevauchants, etc) ne constituent pas une règle générale pour l'ensemble des petits virus, et donc d'autant moins pour l'ensemble des génomes viraux.

Dans ce cadre, la diversité et la richesse apparemment sans fin de la Matière Noire Virale pourrait en réalité témoigner non pas d'un immense réservoir de gènes distincts, mais plutôt de l'existence de portions non utilisées au sein des génomes viraux environnementaux. L'existence d'une telle partie variable et non soumise aux différentes pressions de sélection pourrait même être considérée comme un avantage évolutif sur le long terme pour le virus, puisqu'elle faciliterait les événements de transfert horizontal ou d'apparition de nouvelles fonctions par mutations ponctuelles.

L'analyse de transcriptomes et protéomes viraux devrait permettre, sur des systèmes "simples" comme les virus chimères, de vérifier l'existence et l'activité des gènes prédits, et potentiellement de mieux comprendre l'organisation et la structure des génomes viraux. De même, l'analyse à grande échelle de viromes environnementaux obtenus par l'intermédiaire de projets de grande envergure tels Tara Oceans (Karsenti *et al.*, 2011; Hingamp *et al.*, 2013), ou le Marine Phage Sequencing (http://www.broadinstitute.org/annotation/viral/Phage/) devrait permettre de mieux comprendre cette Matière Noire Virale, et notamment de détecter au sein de cet ensemble de gènes prédits les séquences conservées entre différents échantillons, et donc potentiellement codantes et fonctionnelles. Plus généralement, de tels projets devraient apporter des informations décisives quant à la distribution de gènes, fonctions et génotypes viraux dans l'environnement.

L'origine, la véritable nature et la diversité du monde viral sont donc encore largement à caractériser. Les approches de métagénomique sont en ce sens pleines de promesses et apporteront vraisemblablement à la fois un ensemble de réponses, et sans nul doute un ensemble au moins aussi important de nouvelles questions dans les années à venir. Les différentes analyses de métagénomes viraux ont permis de mieux caractériser la diversité des communautés virales autour du globe, la composition de leurs génomes et leurs différents liens évolutifs, mais également de lever le voile sur des pans entiers de la virosphère ignorés jusqu'alors et potentiellement importants, que ce soit pour des questions de santé ou comme acteurs des écosystèmes. Ces approches seront ainsi certainement utilisées dans un panel de disciplines comme l'épidémiologie, l'écologie et la biotechnologie, pour lesquelles la prise en compte des virus semble importante mais difficile à l'heure actuelle. Une meilleure compréhension des communautés virales et des mécanismes impliqués dans leur évolution temporelle et spatiale serait ainsi bénéfique pour la biologie toute entière.

Références Bibliographiques

A

Abrescia N. G. A., Bamford D. H., Grimes J. M., & Stuart D. I. (2012). Structure unifies the viral universe. *Annual review of biochemistry*, *81*, 795–822.

Ackermann H.-W. (2007). 5500 Phages examined in the electron microscope. *Archives of virology*, *152*(2), 227–43.

Ackermann H.-W., & Prangishvili D. (2012). Prokaryote viruses studied by electron microscopy. *Archives of virology*, *157*(10), 1843–9.

Allen L. Z., Ishoey T., Novotny M. A., McLean J. S., Lasken R. S., & Williamson S. J. (2011). Single virus genomics: a new tool for virus discovery. *PLoS One*, *6*(3), e17722.

Allers E., Moraru C., Duhaime M. B., Beneze E., Solonenko N., Canosa J. B., Amann R., & Sullivan M. B. (2013). Single-cell and population level viral infection dynamics revealed by phageFISH, a method to visualize intracellular and free viruses. *Environmental Microbiology*.

Anderson R. E., Brazelton W. J., & Baross J. a. (2011). Using CRISPRs as a metagenomic tool to identify microbial hosts of a diffuse flow hydrothermal vent viral assemblage. *FEMS microbiology ecology*, *77*(1), 120–33.

Angly F. E., Felts B., Breitbart M., Salamon P., Edwards R. A., Carlson C., Chan A. M., Haynes M., Kelley S., Liu H., Mahaffy J. M., Mueller J. E., Nulton J., Olson R., Parsons R., Rayhawk S., Suttle C. A., & Rohwer F. (2006). The marine viromes of four oceanic regions. *PLoS biology*, *4*(11), e368.

Angly F. E., Willner D., Prieto-Davó A., Edwards R. A., Schmieder R., Vega-Thurber R., Antonopoulos D. A., Barott K., Cottrell M. T., Desnues C., Dinsdale E. A., Furlan M., Haynes M., Henn M. R., Hu Y., Kirchman D. L., McDole T., McPherson J. D., Meyer F., Miller R. M., Mundt E., Naviaux R. K., Rodriguez-Mueller B., Stevens R., Wegley L., Zhang L., Zhu B., & Rohwer F. (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS computational biology*, *5*(12), e1000593.

Angly F., Rodriguez-Brito B., Bangor D., McNairnie P., Breitbart M., Salamon P., Felts B., Nulton J., Mahaffy J., & Rohwer F. (2005). PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics*, *6*, 41.

Arslan D., Legendre M., Seltzer V., Abergel C., & Claverie J.-M. (2011). Distant Mimivirus relative with a larger genome highlights the fundamental features of Megaviridae. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(42), 17486–91.

Arumugam M., Raes J., Pelletier E., Le Paslier D., Yamada T., Mende D. R., Fernandes G. R., Tap J., Bruls T., Batto J.-M., Bertalan M., Borruel N., Casellas F., Fernandez L., Gautier L., Hansen T., Hattori M., Hayashi T., Kleerebezem M., Kurokawa K., Leclerc M., *et al.* (2011). Enterotypes of the human gut microbiome. *Nature*, *473*(7346), 174–80.

Atanasova N. S., Roine E., Oren A., Bamford D. H., & Oksanen H. M. (2012). Global network of specific virus-host interactions in hypersaline environments. *Environmental microbiology*, *14*(2), 426–40.

Aw T. G., & Rose J. B. (2012). Detection of pathogens in water: from phylochips to qPCR to pyrosequencing. *Current opinion in biotechnology*, *23*(3), 422–30.

B

Baltimore D. (1971). Expression of animal virus genomes. *Bacteriological reviews*, *35*(3), 235–41.

Bamford D. H., Grimes J. M., & Stuart D. I. (2005). What does structure tell us about virus evolution? *Current opinion in structural biology*, *15*(6), 655–63.

Barzon L., Lavezzo E., Costanzi G., Franchin E., Toppo S., & Palù G. (2013). Next-generation sequencing technologies in diagnostic virology. *Journal of clinical virology*, 3–7.

Bath C., Cukalac T., Porter K., & Dyall-Smith M. L. (2006). His1 and His2 are distantly related, spindle-shaped haloviruses belonging to the novel virus group, Salterprovirus. *Virology*, *350*(1), 228–39.

Bath C., & Dyall-Smith M. L. (1998). His1, an archaeal virus of the Fuselloviridae family that infects Haloarcula hispanica. *Journal of virology*, *72*(11), 9392–5.

Bawden F. C. (1941). *Plant viruses and virus diseases*. *Chronica botanica* (Chronica B., pp. 248–257). Waltham, MA.

Bell P. J. L. (2009). The viral eukaryogenesis hypothesis: a key role for viruses in the emergence of eukaryotes from a prokaryotic world environment. *Annals of the New York Academy of Sciences*, *1178*, 91–105.

Benson S. D., Bamford J. K. H., Bamford D. H., & Burnett R. M. (2004). Does common architecture reveal a viral lineage spanning all three domains of life? *Molecular cell*, *16*(5), 673–85.

Bergeron A., Belcaid M., Steward G. F., & Poisson G. (2007). Divide and conquer: enriching environmental sequencing data. *PLoS One*, *2*(9), e830.

Bergh O., Børsheim K. Y., Bratbak G., & Heldal M. (1989). High abundance of viruses found in aquatic environments. *Nature*, *340*(6233), 467–468.

Bernal R. A., Hafenstein S., Olson N. H., Bowman V. D., Chipman P. R., Baker T. S., Fane B. A., & Rossmann M. G. (2003). Structural Studies of Bacteriophage α3 Assembly. *Journal of Molecular Biology*, *325*(1), 11–24.

Bettarel Y., Amblard C., Sime-Ngando T., Carrias J.-F., Sargos D., Garabétian F., & Lavandier P. (2003a). Viral lysis, flagellate grazing potential, and bacterial production in Lake Pavin. *Microbial ecology*, *45*(2), 119–27.

Bettarel Y., Bouvier T., Bouvier C., Carré C., Desnues A., Domaizon I., Jacquet S., Robin A., & Sime-Ngando T. (2011). Ecological traits of planktonic viruses and prokaryotes along a full-salinity gradient. *FEMS microbiology ecology*, *76*(2), 360–72.

Bettarel Y., Sime-Ngando T., Amblard C., Carrias J.-F., & Portelli C. (2003b). Virioplankton and microbial communities in aquatic systems: a seasonal study in two lakes of differing trophy. *Freshwater Biology*, *48*(5), 810–822.

Bettarel Y., Sime-Ngando T., Amblard C., & Dolan J. (2004). Viral activity in two contrasting lake ecosystems. *Applied and environmental microbiology*, *70*(5), 2941–2951.

Bibby K. (2013). Metagenomic identification of viral pathogens. *Trends in biotechnology*, *31*(5), 275–279.

Blinkova O., Victoria J., Li Y., Keele B. F., Sanz C., Ndjango J.-B. N., Peeters M., Travis D., Lonsdorf E. V, Wilson M. L., Pusey A. E., Hahn B. H., & Delwart E. L. (2010). Novel circular DNA viruses in stool samples of wild-living chimpanzees. *The Journal of general virology*, *91*, 74–86.

Boucher D., Jardillier L., & Debroas D. (2006). Succession of bacterial community composition over two consecutive years in two aquatic systems: a natural lake and a lake-reservoir. *FEMS microbiology ecology*, *55*(1), 79–97.

Bråte J., Logares R., Berney C., Ree D. K., Klaveness D., Jakobsen K. S., & Shalchian-Tabrizi K. (2010). Freshwater Perkinsea and marine-freshwater colonizations revealed by pyrosequencing and phylogeny of environmental rDNA. *The ISME journal*, *4*(9), 1144–53.

Breitbart M., Felts B., Kelley S., Mahaffy J. M., Nulton J., Salamon P., & Rohwer F. (2004a). Diversity and population structure of a near-shore marine-sediment viral community. *Proceedings. Biological sciences / The Royal Society*, *271*(1539), 565–574.

Breitbart M., Hewson I., Felts B., Mahaffy J. M., Nulton J., Salamon P., & Rohwer F. (2003). Metagenomic Analyses of an Uncultured Viral Community from Human Feces. *Journal of bacteriology*, *185*(20), 6220–6223.

Breitbart M., Miyake J. H., & Rohwer F. (2004b). Global distribution of nearly identical phage-encoded DNA sequences. *FEMS microbiology letters*, *236*(2), 249–256.

Breitbart M., & Rohwer F. (2005). Here a virus, there a virus, everywhere the same virus? *Trends in microbiology*, *13*(6), 278–284.

Breitbart M., Salamon P., Andresen B., Mahaffy J. M., Segall A. M., Mead D., Azam F., & Rohwer F. (2002). Genomic analysis of uncultured marine viral communities. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(22), 14250–5.

Brentlinger K. L., Hafenstein S., Novak C. R., Fane B. A., Borgon R., Mckenna R., & Agbandje-McKenna M. (2002). Microviridae : a Family Divided: Isolation, Characterization, and Genome Sequence of PhiMH2K, a Bacteriophage of the Obligate Intracellular Parasitic Bacterium. *Journal of bacteriology*, *184*(4), 1089–1094.

Brown B. E. (1997). Coral bleaching: causes and consequences. *Coral Reefs*, *16*, 129–138.

Brum J., Culley A., & Steward G. (2013). Assembly of a Marine Viral Metagenome after Physical Fractionation. *PLoS One*, *8*(4), e60604.

Brüssow H. (2009). The not so universal tree of life or the place of viruses in the living world. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *364*(1527), 2263–74.

Busby B., Kristensen D. M., & Koonin E. V. (2013). Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens. *Environmental microbiology*, *15*(2), 307–312.

C

Canchaya C., Fournous G., & Brüssow H. (2004). The impact of prophages on bacterial chromosomes. *Molecular Microbiology*, *53*(1), 9–18.

Carrillo C., Tulman E. R., Delhon G., Lu Z., Carreno A., Kutish G. F., & Rock D. L. (2005). Comparative Genomics of Foot-and-Mouth Disease Virus. *Journal of virology*, *79*(10), 6487–6504.

Carrillo-Tripp M., Shepherd C. M., Borelli I. a, Venkataraman S., Lander G., Natarajan P., Johnson J. E., Brooks C. L., & Reddy V. S. (2009). VIPERdb2: an enhanced and web API enabled relational database for structural virology. *Nucleic acids research*, *37*(Database issue), D436–42.

Casamayor E. O., Massana R., Benlloch S., Øvreås L., Díez B., Goddard V. J., Gasol J. M., Joint I., Rodríguez-Valera F., & Pedrós-Alió C. (2002). Changes in archaeal, bacterial and eukaryal assemblages along a salinity gradient by comparison of genetic fingerprinting methods in a multipond solar saltern. *Environmental microbiology*, *4*(6), 338–48.

Casas V., & Rohwer F. (2007). Phage metagenomics. *Methods in enzymology*, *421*, 259–68.

Case R. J., & Boucher Y. (2011). Molecular musings in microbial ecology and evolution. *Biology direct*, *6*(1), 58.

Chastel C. (1992). *Histoire des virus : de la variole au sida* (Boubée.). Paris.

Chastel C. (1997). La naissance de la virologie. *Virologie*, *1*(2), 103–110.

Cherwa J., & Fane B. A. (2011). Microviridae: microviruses and gokushoviruses. In *eLS* (John Wiley.).

Clasen J. L., & Suttle C. A. (2009). Identification of freshwater Phycodnaviridae and their potential phytoplankton hosts, using DNA pol sequence fragments and a genetic-distance analysis. *Applied and environmental microbiology*, *75*(4), 991–997.

Claverie J.-M. (2006). Viruses take center stage in cellular evolution. *Genome biology*, *7*(6), 110.

Claverie J.-M., & Abergel C. (2013). *Open questions about giant viruses. Advances in virus research* (1st ed., Vol. 85, pp. 25–56). Elsevier Inc.

Clerissi C., Desdevises Y., & Grimsley N. (2012). Prasinoviruses of the marine green alga Ostreococcus tauri are mainly species specific. *Journal of virology*, *86*(8), 4611–9.

Colombet J. (2008). *Importance de la variabilité verticale dans un lac méromictique profond : diversité et activité lysogène des communautés virales*. Thèse : doctorat. Université Blaise Pascal, Clermont-Ferrand

Colombet J., Sime-Ngando T., Cauchie H. M., Fonty G., Hoffmann L., & Demeure G. (2006). Depth-related gradients of viral activity in Lake Pavin. *Applied and environmental microbiology*, *72*(6), 4440–5.

Comeau A. M., Arbiol C., & Krisch H. M. (2010). Gene network visualization and quantitative synteny analysis of more than 300 marine T4-like phage scaffolds from the GOS metagenome. *Molecular biology and evolution*, *27*(8), 1935–44.

Comeau A. M., Tremblay D., Moineau S., Rattei T., Kushkina A. I., Tovkach F. I., Krisch H. M., & Ackermann H.-W. (2012). Phage morphology recapitulates phylogeny: the comparative genomics of a new group of myoviruses. *PLoS One*, *7*(7), e40102.

Correa A. M. S., Welsh R. M., & Thurber R. L. V. (2013). Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *The ISME journal*, *7*(1), 13–27.

Cresawn S. G., Bogel M., Day N., Jacobs-Sera D., Hendrix R. W., & Hatfull G. F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC bioinformatics*, *12*, 395.

Culley A. I., Lang A. S., & Suttle C. a. (2006). Metagenomic analysis of coastal RNA virus communities. *Science*, *312*(5781), 1795–8.

D

Danna K. J., Sack Jr G. H., & Nathans D. (1973). Studies of Simian virus 40 DNA: VII. A cleavage map of the SV40 genome. *Journal of Molecular Biology*, *78*(2), 363–376.

De Villiers E. P., Gallardo C., Arias M., da Silva M., Upton C., Martin R., & Bishop R. P. (2010). Phylogenomic analysis of 11 complete African swine fever virus genome sequences. *Virology*, *400*(1), 128–36.

De Wit R., & Bouvier T. (2006). “Everything is everywhere, but, the environment selects”; what did Baas Becking and Beijerinck really say? *Environmental microbiology*, *8*(4), 755–8.

Delwart E., & Li L. (2012). Rapidly expanding genetic diversity and host range of the Circoviridae viral family and other Rep encoding small circular ssDNA genomes. *Virus research*, *164*(1-2), 114–21.

Deng L., Gregory A., Yilmaz S., & Poulos B. (2012). Contrasting life strategies of viruses that infect photo-and heterotrophic bacteria, as revealed by viral tagging. *Mbio*, *3*(6), e00373–12.

Desnues C., & Raoult D. (2012). Virophages question the existence of satellites. *Nature reviews. Microbiology*, *10*(3), 234; author reply 234.

Desnues C., Rodriguez-Brito B., Rayhawk S., Kelley S., Tran T., Haynes M., Liu H., Furlan M., Wegley L., Chau B., Ruan Y., Hall D., Angly F. E., Edwards R. a, Li L., Thurber R. V., Reid R. P., Siefert J., Souza V., Valentine D. L., Swan B. K., Breitbart M., & Rohwer F. (2008). Biodiversity and biogeography of phages in modern stromatolites and thrombolites. *Nature*, *452*(7185), 340–3.

Diemer G. S., & Stedman K. M. (2012). A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biology direct*, *7*, 13.

Dinsdale E. A., Edwards R. A., Hall D., Angly F., Breitbart M., Brulc J. M., Furlan M., Desnues C., Haynes M., Li L., McDaniel L., Moran M. A., Nelson K. E., Nilsson C., Olson R., Paul J., Brito B. R., Ruan Y., Swan B. K., Stevens R., Valentine D. L., Thurber R. V., Wegley L., White B. A., & Rohwer F. (2008). Functional metagenomic profiling of nine biomes. *Nature*, *452*(7187), 629–32.

Djikeng A., Kuzmickas R., Anderson N. G., & Spiro D. J. (2009). Metagenomic analysis of RNA viruses in a fresh water lake. *PLoS One*, *4*(9), e7264.

Dorigo U., Fontvieille D., & Humbert J.-F. (2006). Spatial variability in the abundance and composition of the free-living bacterioplankton community in the pelagic zone of Lake Bourget (France). *FEMS microbiology ecology*, *58*(1), 109–119.

Douglas a E. (2003). Coral bleaching--how and why? *Marine pollution bulletin*, *46*(4), 385–92.

Duckworth D. H. (1976). Who discovered bacteriophage? *Bacteriological reviews*, *40*(4), 793–802.

Duffy S., Shackelton L. a, & Holmes E. C. (2008). Rates of evolutionary change in viruses: patterns and determinants. *Nature reviews. Genetics*, *9*(4), 267–76.

Duhaime M. B., Deng L., Poulos B. T., & Sullivan M. B. (2012). Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environmental microbiology*, *14*(9), 2526–37.

Duhaime M. B., & Sullivan M. B. (2012). Ocean viruses: Rigorously evaluating the metagenomic sample-to-sequence pipeline. *Virology*, *434*(2), 181–186.

E

Ebert D., & Bull J. J. (2007). The evolution and expression of virulence. In S. C. Stearns & J. C. Koella (Eds.), *Evolution in Health and Disease* (Oxford Uni., pp. 153–167). Oxford.

Eddy S. R. (2011). Accelerated Profile HMM Searches. *PLoS computational biology*, *7*(10), e1002195.

Edwards R. A., & Rohwer F. (2005). Viral metagenomics. *Nature Reviews Microbiology*, *3*(6), 504–510.

Emerson J. B., Thomas B. C., Andrade K., Allen E. E., Heidelberg K. B., & Banfield J. F. (2012). Metagenomic assembly reveals dynamic viral populations in hypersaline systems. *Applied and environmental microbiology*, *78*(17), 6309 – 6320.

F

Fancello L., Raoult D., & Desnues C. (2012a). Computational tools for viral metagenomics and their application in clinical research. *Virology*, *434*(2), 162–174.

Fancello L., Trape S., Robert C., Boyer M., Popgeorgiev N., Raoult D., & Desnues C. (2012b). Viruses in the desert: a metagenomic survey of viral communities in four perennial ponds of the Mauritanian Sahara. *The ISME journal*, *7*, 1–11.

Faruque S. M., & Mekalanos J. J. (2003). Pathogenicity islands and phages in Vibrio cholerae evolution. *Trends in Microbiology*, *11*(11), 505–510.

Fauquet C. M., & Fargette D. (2005). International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virology journal*, *2*, 64.

Fenchel T., & Finlay B. J. (2004). The Ubiquity of Small Species: Patterns of Local and Global Diversity. *BioScience*, *54*(8), 777.

Fiers W., Contreras R., Duerinck F., Haegeman G., Iserentant D., Merregaert J., Min Jou W., Molemans F., Raeymaekers A., Van den Berghe A., Volckaert G., & Ysebaert M. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA : primary and secondary structure of the replicase gene. *Nature*, *260*.

Finkbeiner S. R., Allred A. F., Tarr P. I., Klein E. J., Kirkwood C. D., & Wang D. (2008). Metagenomic analysis of human diarrhea: viral detection and discovery. *PLoS pathogens*, *4*(2), e1000011.

Fischer M. G. (2012). Sputnik and Mavirus: more than just satellite viruses. *Nature reviews. Microbiology*, *10*(1), 78; author reply 78.

Forterre P. (2006). The origin of viruses and their possible roles in major evolutionary transitions. *Virus research*, *117*(1), 5–16.

Forterre P. (2013). The virocell concept and environmental microbiology. *The ISME journal*, *7*(2), 233–6.

Forterre P., & Prangishvili D. (2009). The great billion-year war between ribosome- and capsid-encoding organisms (cells and viruses) as the major source of evolutionary novelties. *Annals of the New York Academy of Sciences*, *1178*, 65–77.

Forterre P., Soler N., Krupovic M., Marguet E., & Ackermann H.-W. (2013). Fake virus particles generated by fluorescence microscopy. *Trends in Microbiology*, *21*(1), 1–5.

Fraenkel-Conrat H., & Williams R. C. (1955). Reconstitution of active tobacco mosaic virus from inactive protein and nucleic acid components. *Proceedings of the National Academy of Sciences of the United States of America*, *41*(10), 690–698.

Franklin R. E. (1955). Structure of Tobacco Mosaic Virus. *Nature*, *175*, 379–381.

Fuhrman J. A. (1999). Marine viruses and their biogeochemical and ecological effects. *Nature*, *399*(6736), 541–548.

G

Garcia-Heredia I., Martin-Cuadrado A.-B., Mojica F. J. M., Santos F., Mira A., Antón J., & Rodriguez-Valera F. (2012). Reconstructing Viral Genomes from the Environment Using Fosmid Clones: The Case of Haloviruses. *PLoS One*, *7*(3), e33802.

Ghosh T. S., Mohammed M. H., Komanduri D., & Mande S. S. (2011). ProViDE: A software tool for accurate estimation of viral diversity in metagenomic samples. *Bioinformation*, *6*(2), 91–4.

Grayson P., & Molineux I. J. (2007). Is phage DNA "injected" into cells -biologists and physicists can agree. *Current opinion in microbiology*, *10*(4), 401–9.

Greninger A. L., Chen E. C., Sittler T., Scheinerman A., Roubinian N., Yu G., Kim E., Pillai D. R., Guyard C., Mazzulli T., Isa P., Arias C. F., Hackett J., Schochetman G., Miller S., Tang P., & Chiu C. Y. (2010). A metagenomic analysis of pandemic influenza A (2009 H1N1) infection in patients from North America. *PLoS One*, *5*(10), e13381.

H

Handelsman J., Rondon M. R., Brady S. F., Clardy J., & Goodman R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & biology*, *5*(10), R245–9.

Häring M., Rachel R., Peng X., Garrett R. A., & Prangishvili D. (2005). Viral Diversity in Hot Springs of Pozzuoli, Italy, and Characterization of a Unique Archaeal Virus, Acidianus Bottle-Shaped Virus, from a New Family, the Ampullaviridae. *Journal of virology*, *79*(15), 9904–9911.

Held N. L., & Whitaker R. J. (2009). Viral biogeography revealed by signatures in Sulfolobus islandicus genomes. *Environmental microbiology*, *11*(2), 457–66.

Helton R. R., & Wommack K. E. (2009). Seasonal dynamics and metagenomic characterization of estuarine viriobenthos assemblages by randomly amplified polymorphic DNA PCR. *Applied and environmental microbiology*, *75*(8), 2259–65.

Hendrix R. W., Lawrence J. G., Hatfull G. F., & Casjens S. (2000). The origins and ongoing evolution of viruses. *Trends in microbiology*, *8*(11), 504–8.

Hingamp P., Grimsley N., Acinas S. G., Clerissi C., Subirana L., Poulain J., Ferrera I., Sarmento H., Villar E., Lima-Mendez G., Faust K., Sunagawa S., Claverie J.-M., Moreau H., Desdevises Y., Bork P., Raes J., de Vargas C., Karsenti E., Kandels-Lewis S., Jaillon O., Not F., Pesant S., Wincker P., & Ogata H. (2013). Exploring nucleo-

cytoplasmic large DNA viruses in Tara Oceans microbial metagenomes. *The ISME journal*, 1–18.

Hoyle L. (1952). Structure of the influenza virus. The relation between biological activity and chemical structure of virus fractions. *Journal of Hygiène*, *50*(2), 229–245.

Hurwitz B. L., Deng L., Poulos B. T., & Sullivan M. B. (2012). Evaluation of methods to concentrate and purify ocean virus communities through comparative, replicated metagenomics. *Environmental Microbiology*. *15,* 1428-1440.

Hurwitz B. L., & Sullivan M. B. (2013). The Pacific Ocean Virome (POV): a marine viral metagenomic dataset and associated protein clusters for quantitative viral ecology. *PLoS One*, *8*(2), e57355.

Huttenhower C., & Hofmann O. (2010). A quick guide to large-scale genomic data mining. *PLoS computational biology*, *6*(5), e1000779.

I

Ibiricu I., Huiskonen J., Döhner K., & Bradke F. (2011). Cryo electron tomography of herpes simplex virus during axonal transport and secondary envelopment in primary neurons. *PLoS pathogens*, *7*(12), e1002406.

Ignacio-Espinoza J. C., & Sullivan M. B. (2012). Phylogenomics of T4 cyanophages: lateral gene transfer in the "core" and origins of host genes. *Environmental microbiology*.

Ilyas M., Qazi J., Mansoor S., & Briddon R. W. (2009). Molecular characterisation and infectivity of a "Legumovirus" (genus Begomovirus: family Geminiviridae) infecting the leguminous weed Rhynchosia minima in Pakistan. *Virus Research*, *145*(2), 279–284.

J

Jacquet S., Domaizon I., Personnic S., Pradeep Ram A. S., Hedal M., Duhamel S., & Sime-Ngando T. (2005). Estimates of protozoan- and viral-mediated mortality of bacterioplankton in Lake Bourget (France). *Freshwater Biology*, *50* , 627–645.

Jaschke P. R., Lieberman E. K., Rodriguez J., Sierra A., & Endy D. (2012). A fully decompressed synthetic bacteriophage øX174 genome assembled and archived in yeast. *Virology*, *434*(2), 278–84.

Johnston I. G., Louis A. A., & Doye J. P. K. (2010). Modelling the self-assembly of virus capsids. *Journal of physics. Condensed matter : an Institute of Physics journal*, *22*(10), 104101.

Julien J.-P., Lee P. S., & Wilson I. a. (2012). Structural insights into key sites of vulnerability on HIV-1 Env and influenza HA. *Immunological reviews*, *250*(1), 180–98.

K

Karsenti E., Acinas S. G., Bork P., Bowler C., De Vargas C., Raes J., Sullivan M., Arendt D., Benzoni F., Claverie J.-M., Follows M., Gorsky G., Hingamp P., Iudicone D., Jaillon O.,

Kandels-Lewis S., Krzic U., Not F., Ogata H., Pesant S., Reynaud E. G., Sardet C., Sieracki M. E., Speich S., Velayoudon D., Weissenbach J., & Wincker P. (2011). A holistic approach to marine eco-systems biology. *PLoS biology*, *9*(10), e1001177.

Kim M.-S., Park E.-J., Roh S. W., & Bae J.-W. (2011). Diversity and abundance of single-stranded DNA viruses in human feces. *Applied and environmental microbiology*, *77*(22), 8062–8070.

Klein R., Baranyi U., Greineder B., Scholz H., & Witte A. (2002). Natrialba magadii virus PhiCh1 : first complete nucleotide sequence and functional organization of a virus infecting a haloalkaliphilic archaeon. *Molecular Microbiology*, *45*(3), 851–863.

Klingenberg H., Petra Aßhauer K., Lingner T., & Meinicke P. (2013). Protein signature-based estimation of metagenomic abundances including all domains of life and viruses. *Bioinformatics*, 1–8.

Knipe D. M., & Cliffe A. (2008). Chromatin control of herpes simplex virus lytic and latent infection. *Nature reviews. Microbiology*, *6*(3), 211–21.

Koonin E. V, Senkevich T. G., & Dolja V. V. (2006). The ancient Virus World and evolution of cells. *Biology direct*, *1*, 29.

Kristensen D. M., Cai X., & Mushegian A. (2011). Evolutionarily conserved orthologous families in phages are relatively rare in their prokaryotic hosts. *Journal of bacteriology*, *193*(8), 1806–14.

Kristensen D. M., Mushegian A. R., Dolja V. V, & Koonin E. V. (2010). New dimensions of the virus world discovered through metagenomics. *Trends in microbiology*, *18*(1), 11–19.

Kristensen D. M., Waller A. S., Yamada T., Bork P., Mushegian A. R., & Koonin E. V. (2013). Orthologous Gene Clusters and Taxon Signature Genes for Viruses of Prokaryotes. *Journal of Bacteriology*, *195*(5), 941–950.

Krupovic M., & Bamford D. H. (2009). Does the evolution of viral polymerases reflect the origin and evolution of viruses? *Nature reviews. Microbiology*, *7*(3), 250; author reply 250.

Krupovic M., & Cvirkaite-Krupovic V. (2011). Virophages or satellite viruses? *Nature Reviews Microbiology*, *9*(11), 762–763.

Krupovic M., & Forterre P. (2011). Microviridae goes temperate: microvirus-related proviruses reside in the genomes of Bacteroidetes. *PLoS One*, *6*(5), e19893.

Krupovic M., Ravantti J. J., & Bamford D. H. (2009). Geminiviruses: a tale of a plasmid becoming a virus. *BMC evolutionary biology*, *9*, 112.

Kunin V., He S., Warnecke F., Peterson S. B., Garcia Martin H., Haynes M., Ivanova N., Blackall L. L., Breitbart M., Rohwer F., McMahon K. D., & Hugenholtz P. (2008). A bacterial metapopulation adapts locally to phage predation despite global dispersal. *Genome research*, *18*(2), 293–7.

Kuznetsov Y., Gershon P. D., & McPherson A. (2008). Atomic force microscopy investigation of vaccinia virus structure. *Journal of virology*, *82*(15), 7551–66.

L

Lee H. S., & Sobsey M. D. (2011). Survival of prototype strains of somatic coliphage families in environmental waters and when exposed to UV low-pressure monochromatic radiation or heat. *Water Research*, *45*(12), 3723–3734.

Lefeuvre P., Harkins G. W., Lett J.-M., Briddon R. W., Chase M. W., Moury B., & Martin D. P. (2011). Evolutionary time-scale of the begomoviruses: evidence from integrated sequences in the Nicotiana genome. *PLoS One*, *6*(5), e19193.

Legendre M., Santini S., Rico A., Abergel C., & Claverie J.-M. (2011). Breaking the 1000-gene barrier for Mimivirus using ultra-deep genome and transcriptome sequencing. *Virology journal*, *8*(1), 99.

Leplae R., Lima-Mendez G., & Toussaint A. (2010). ACLAME: a CLAssification of Mobile genetic Elements, update 2010. *Nucleic acids research*, *38*(Database issue), D57–61.

Li W., Zhang T., Tang X., & Wang B. (2010). Oomycetes and fungi: important parasites on marine algae. *Acta Oceanologica Sinica*, *29*(5), 74–81.

Liao Y.-C., Lee M.-S., Ko C.-Y., & Hsiung C. a. (2008). Bioinformatics models for predicting antigenic variants of influenza A/H3N2 virus. *Bioinformatics*, *24*(4), 505–12.

Lima-Mendez G., Van Helden J., Toussaint A., & Leplae R. (2008). Reticulate representation of evolutionary and functional relationships between phage genomes. *Molecular biology and evolution*, *25*(4), 762–77.

Lin L., Bitner R., & Edlin G. (1977). Increased reproductive fitness of Escherichia coli lambda lysogens. *Journal of virology*, *21*(2), 554–9.

Lindell D., Jaffe J. D., Coleman M. L., Futschik M. E., Axmann I. M., Rector T., Kettler G., Sullivan M. B., Steen R., Hess W. R., Church G. M., & Chisholm S. W. (2007). Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution. *Nature*, *449*(7158), 83–6.

Lindell D., Jaffe J. D., Johnson Z. I., Church G. M., & Chisholm S. W. (2005). Photosynthesis genes in marine viruses yield proteins during host infection. *Nature*, *438*(7064), 86–9.

Lindell D., Sullivan M. B., Johnson Z. I., Tolonen A. C., Rohwer F., & Chisholm S. W. (2004). Transfer of photosynthesis genes to and from Prochlorococcus viruses. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(30), 11013–8.

Liu H., Fu Y., Li B., Yu X., Xie J., Cheng J., Ghabrial S. a, Li G., Yi X., & Jiang D. (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC evolutionary biology*, *11*, 276.

Liu J., Glazko G., & Mushegian A. (2006). Protein repertoire of double-stranded DNA bacteriophages. *Virus research*, *117*, 68–80.

Logares R., Bråte J., Bertilsson S., Clasen J. L., Shalchian-Tabrizi K., & Rengefors K. (2009). Infrequent marine-freshwater transitions in the microbial world. *Trends in microbiology*, *17*(9), 414–422.

Logares R., Lindström E. S., Langenheder S., Logue J. B., Paterson H., Laybourn-Parry J., Rengefors K., Tranvik L., & Bertilsson S. (2012). Biogeography of bacterial communities exposed to progressive long-term environmental change. *The ISME journal*, 937–948.

Lopes C. T., Franz M., Kazi F., Donaldson S. L., Morris Q., & Bader G. D. (2010). Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, *26*(18), 2347–8.

López-Bueno A., Tamames J., Velázquez D., Moya A., Quesada A., & Alcamí A. (2009). High diversity of the viral community from an Antarctic lake. *Science*, *326*(5954), 858–61.

Lopez-Garcia P. (2012). The Place of Viruses in Biology in Light of the Metabolism- versus-replication-first Debate. *History and Philosophy of the Life Sciences*, *34*, 391–406.

Lorenzi H. a, Hoover J., Inman J., Safford T., Murphy S., Kagan L., & Williamson S. J. (2011). TheViral MetaGenome Annotation Pipeline(VMGAP):an automated tool for the functional annotation of viral Metagenomic shotgun sequencing data. *Standards in genomic sciences*, *4*(3), 418–29.

Lwoff A. (1957). The Concept of Virus. *Journal of General Microbiology*, *17*(1), 239–253.

M

Mahy B. W. J. (2005). Introduction and history of foot-and-mouth disease virus. *Current topics in microbiology and immunology*, *288*, 1–8.

Mansoor S., Qazi J., Amin I., Khatri A., Khan I. A., Raza S., Zafar Y., & Briddon R. W. (2005). A PCR-Based Method, With Internal Control, for the Detection of Banana Bunchy Top Virus in Banana. *Molecular Biotechnology*, *30*, 167–169.

Marhaver K. L., Edwards R. A., & Rohwer F. (2008). Viral communities associated with healthy and bleaching corals. *Environmental microbiology*, *10*(9), 2277–2286.

Márquez L. M., Redman R. S., Rodriguez R. J., & Roossinck M. J. (2007). A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. *Science*, *315*(5811), 513–5.

Marshall J. A. (2012). The role of transmission electron microscopy in the study of gastroenteritis viruses. *Microbiology Australia*, 85–86.

Martín-Cuadrado A.-B., López-García P., Alba J.-C., Moreira D., Monticelli L., Strittmatter A., Gottschalk G., & Rodríguez-Valera F. (2007). Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS One*, *2*(9), e914.

Marvin Seibert M., Ekeberg T., Maia F. R. N. C., Svenda M., Andreasson J., Jönsson O., Odić D., Iwan B., Rocker A., Westphal D., Hantke M., DePonte D. P., Barty A., Schulz J., Gumprecht L., Coppola N., Aquila A., Liang M., White T. a, Martin A., Caleman C.,

et al. (2011). Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*, *470*(7332), 78–81.

Mathan M., Swaminathan S. P., Mathan V. I., Yesudoss S., & Baker S. J. (1975). Pleomorphic Virus-like Particles in Human Faeces. *The Lancet*, *305*(7915), 1068–1069.

McDaniel L., Breitbart M., Mobberley J., Long A., Haynes M., Rohwer F., & Paul J. H. (2008). Metagenomic analysis of lysogeny in Tampa Bay: implications for prophage gene expression. *PLoS One*, *3*(9), e3263.

McDaniel L. D., Rosario K., Breitbart M., & Paul J. H. (2013). Comparative metagenomics: Natural populations of induced prophages demonstrate highly unique, lower diversity viral sequences. *Environmental Microbiology*,

Merckel M. C., Huiskonen J. T., Bamford D. H., Goldman A., & Tuma R. (2005). The structure of the bacteriophage PRD1 spike sheds light on the evolution of viral capsid architecture. *Molecular cell*, *18*(2), 161–70.

Meyer F., Paarmann D., D'Souza M., Olson R., Glass E. M., Kubal M., Paczian T., Rodriguez A., Stevens R., Wilke A., Wilkening J., & Edwards R. a. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC bioinformatics*, *9*, 386.

Meyer J. R., Dobias D. T., Weitz J. S., Barrick J. E., Quick R. T., & Lenski R. E. (2012). Repeatability and contingency in the evolution of a key innovation in phage lambda. *Science*, *335*(6067), 428–32.

Minot S., Grunberg S., Wu G. D., Lewis J. D., & Bushman F. D. (2012a). Hypervariable loci in the human gut virome. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(10), 3962–6.

Minot S., Sinha R., Chen J., Li H., Keilbaugh S. A., Wu G. D., Lewis J. D., & Bushman F. D. (2011). The human gut virome: inter-individual variation and dynamic response to diet. *Genome Research*, *21*(10), 1616–1625.

Minot S., Wu G. D., Lewis J. D., & Bushman F. D. (2012b). Conservation of Gene Cassettes among Diverse Viruses of the Human Gut. *PLoS One*, *7*(8), e42342.

Monier A., Claverie J.-M., & Ogata H. (2008). Taxonomic distribution of large DNA viruses in the sea. *Genome biology*, *9*(7), R106.

Monier A., Pagarete A., de Vargas C., Allen M. J., Read B., Claverie J.-M., & Ogata H. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome research*, *19*(8), 1441–9.

Mora C., Tittensor D. P., Adl S., Simpson A. G. B., & Worm B. (2011). How many species are there on Earth and in the ocean? *PLoS biology*, *9*(8), e1001127.

Moran N. A. (2002). Microbial minimalism: genome reduction in bacterial pathogens. *Cell*, *108*(5), 583–6.

Moreau H., Piganeau G., Desdevises Y., Cooke R., Derelle E., & Grimsley N. (2010). Marine prasinovirus genomes show low evolutionary divergence and acquisition of protein metabolism genes by horizontal gene transfer. *Journal of virology*, *84*(24), 12555–63.

Moreira D., & López-García P. (2009). Ten reasons to exclude viruses from the tree of life. *Nature reviews. Microbiology*, *7*(4), 306–11.

N

Nakamura S., Yang C.-S., Sakon N., Ueda M., Tougan T., Yamashita A., Goto N., Takahashi K., Yasunaga T., Ikuta K., Mizutani T., Okamoto Y., Tagami M., Morita R., Maeda N., Kawai J., Hayashizaki Y., Nagai Y., Horii T., Iida T., & Nakaya T. (2009). Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One*, *4*(1), e4219.

Nasir A., Kim K. M., & Caetano-Anolles G. (2012). Giant viruses coexisted with the cellular ancestors and represent a distinct supergroup along with superkingdoms Archaea, Bacteria and Eukarya. *BMC evolutionary biology*, *12*(1), 156.

Ng T. F. F., Marine R., Wang C., Simmonds P., Kapusinszky B., Bodhidatta L., Oderinde B. S., Wommack K. E., & Delwart E. (2012). High variety of known and new RNA and DNA viruses of diverse origins in untreated sewage. *Journal of virology*, *86*(22), 12161–75.

O

Oksanen J., Kindt R., Legendre P., O'Hara B., Simpson G. L., Solymos P., Stevens M. H. H., & Wagner H. (2008). *The vegan Package*.

Ondov B. D., Bergman N. H., & Phillippy A. M. (2011). Interactive metagenomic visualization in a Web browser. *BMC bioinformatics*, *12*(1), 385.

Ouardani M., Wilson L., Jetté R., Montpetit C., & Dea S. (1999). Multiplex PCR for detection and typing of porcine circoviruses. *Journal of clinical microbiology*, *37*(12), 3917–24.

P

Pal C., Maciá M. D., Oliver A., Schachar I., & Buckling A. (2007). Coevolution with viruses drives the evolution of bacterial mutation rates. *Nature*, *450*(7172), 1079–81.

Palacios G., Druce J., Du L., Tran T., Birch C., Briese T., Conlan S., Quan P., Hui J., Marshall J., Simons J. F., Egholm M., Paddock C. D., Shieh W., Goldsmith C. S., Zaki S. R., Catton M., & Lipkin W. I. (2008). A new arenavirus in a cluster of fatal transplant-associated diseases. *The New England Journal of Medicine*, *358*(10), 991–998.

Park E.-J., Kim K.-H., Abell G. C. J., Kim M.-S., Roh S. W., & Bae J.-W. (2011). Metagenomic analysis of the viral communities in fermented foods. *Applied and environmental microbiology*, *77*(4), 1284–91.

Payet J. P., & Suttle C. A. (2013). To kill or not to kill: The balance between lytic and lysogenic viral infection is driven by trophic status. *Limnology and Oceanography*, *58*(2), 465–474.

Pérez-Brocal V., García-López R., Vázquez-Castellanos J. F., Nos P., Beltrán B., Latorre A., & Moya A. (2013). Study of the viral and microbial communities associated with Crohn's disease: a metagenomic approach. *Clinical and translational gastroenterology*, *4*(November 2012), e36.

Pérez-del-Olmo A., Fernández M., Raga J. A., Kostadinova A., & Morand S. (2009). Not everything is everywhere: the distance decay of similarity in a marine host-parasite system. *Journal of Biogeography*, *36*(2), 200–209.

Personnic S., Domaizon I., Dorigo U., Berdjeb L., & Jacquet S. (2009a). Seasonal and spatial variability of virio-, bacterio-, and picophytoplanktonic abundances in three peri-alpine lakes. *Hydrobiologia*, *627, Numbe*, 99–116.

Personnic S., Domaizon I., Sime-Ngando T., & Jacquet S. (2009b). Seasonal variations of microbial abundances and virus- versus flagellate-induced mortality of picoplankton in three peri-alpine lakes. *Journal of Plankton Research*, *31*(10), 1161–1177.

Pommier T., Douzery E. J. P., & Mouillot D. (2012). Environment drives high phylogenetic turnover among oceanic bacterial communities. *Biology letters*, *8*(4), 562–6.

Porter K., Kukkaro P., Bamford J. K. H., Bath C., Kivelä H. M., Dyall-Smith M. L., & Bamford D. H. (2005). SH1: A novel, spherical halovirus isolated from an Australian hypersaline lake. *Virology*, *335*(1), 22–33.

R

Raoult D., & Forterre P. (2008). Redefining viruses: lessons from Mimivirus. *Nature reviews. Microbiology*, *6*(4), 315–9.

Rapaport D. C. (2010). Modeling capsid self-assembly: design and analysis. *Physical biology*, *4*.

Rappé M. S., & Giovannoni S. J. (2003). The uncultured microbial majority. *Annual review of microbiology*, *57*, 369–94.

Reddy V. B., Thimmappaya B., Dhar R., Subramanian K. N., Zain B., Pan J., Ghosh P., Celma M., & Weissman S. (1978). The genome of simian virus 40. *Science*, *200*(4341), 494–502.

Reyes A., Haynes M., Hanson N., Angly F. E., Heath A. C., Rohwer F., & Gordon J. I. (2010). Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature*, *466*(7304), 334–338.

Reyes A., Semenkovich N. P., Whiteson K., Rohwer F., & Gordon J. I. (2012). Going viral: next-generation sequencing applied to phage populations in the human gut. *Nature reviews. Microbiology*, *10*(9), 607–17.

Rodriguez-Brito B., Li L., Wegley L., Furlan M., Angly F., Breitbart M., Buchanan J., Desnues C., Dinsdale E., Edwards R., Felts B., Haynes M., Liu H., Lipson D., Mahaffy J., Martin-Cuadrado A. B., Mira A., Nulton J., Pasić L., Rayhawk S., Rodriguez-Mueller J., Rodriguez-Valera F., Salamon P., Srinagesh S., Thingstad T. F., Tran T., Thurber R. V., Willner D., Youle M., & Rohwer F. (2010). Viral and microbial community dynamics in four aquatic environments. *The ISME journal*, *4*(6), 739–51.

Rodriguez-Valera F., Martin-Cuadrado A.-B., Rodriguez-Brito B., Pasić L., Thingstad T. F., Rohwer F., & Mira A. (2009). Explaining microbial population genomics through phage predation. *Nature Reviews Microbiology*, *7*(11), 828–836.

Rohwer F. (2003). Global phage diversity. *Cell*, *113*(2), 141.

Rohwer F., Segall A., Steward G., Seguritan V., Breitbart M., Wolven F., & Azam F. (2000). The complete genomic sequence of the marine phage Roseophage SIO1 shares homology with nonmarine phages. *Limnology and Oceanography*, *45*(2), 408–418.

Roine E., & Bamford D. H. (2012). Lipids of archaeal viruses. *Archaea*, *2012*, 384919.

Rokyta D. R., Abdo Z., & Wichman H. A. (2009). The genetics of adaptation for eight microvirid bacteriophages. *Journal of molecular evolution*, *69*(3), 229–239.

Rokyta D. R., Burch C. L., Caudle S. B., & Wichman H. A. (2006). Horizontal gene transfer and the evolution of microvirid coliphage genomes. *Journal of bacteriology*, *188*(3), 1134–1142.

Rosario K., Duffy S., & Breitbart M. (2009a). Diverse circovirus-like genome architectures revealed by environmental metagenomics. *Journal of general virology*, *90*(Pt 10), 2418–2424.

Rosario K., Duffy S., & Breitbart M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Archives of virology*, *157*(10), 1851–71.

Rosario K., Nilsson C., Lim Y. W., Ruan Y., & Breitbart M. (2009b). Metagenomic analysis of viruses in reclaimed water. *Environmental microbiology*, *11*(11), 2806–20.

Roux E. (1903). Sur les microbes dits “invisibles.” *Bulletin de l’Institut Pasteur*, *7*, 7–12.

Roux S., Enault F., Bronner G., & Debroas D. (2011). Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in ecosystems. *FEMS microbiology ecology*, *78*(3), 617–28.

Roux S., Enault F., Robin A., Ravet V., Personnic S., Theil S., Colombet J., Sime-Ngando T., & Debroas D. (2012). Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS One*, *7*(3), e33641.

S

Sandaa R., Short S. M., & Schroeder D. C. (2010). Fingerprinting aquatic virus communities. In S. W. Wilhelm, M. G. Weinbauer, & C. A. Suttle (Eds.), *Manual of Aquatic Viral Ecology* (ASLO., pp. 9–18).

Sanger F., Coulson A., Friedmann T., Air G., Barrell B., Brown N., Fiddes J., Hutchison III C., Slocombe P., & Smith M. (1978). The nucleotide sequence of bacteriophage phiX174. *Journal of molecular biology*, *125*, 225–246.

Sano E., Carlson S., Wegley L., & Rohwer F. (2004). Movement of Viruses between Biomes. *Applied and environmental microbiology*, *70*(10), 5842–5846.

Santos F., Yarza P., Parro V., Briones C., & Antón J. (2010). The metavirome of a hypersaline environment. *Environmental microbiology*, *12*(11), 2965–76.

Schloissnig S., Arumugam M., Sunagawa S., Mitreva M., Tap J., Zhu A., Waller A., Mende D. R., Kultima J. R., Martin J., Kota K., Sunyaev S. R., Weinstock G. M., & Bork P. (2013). Genomic variation landscape of the human gut microbiome. *Nature*, *493*(7430), 45–50.

Schoenfeld T., Liles M., Wommack K. E., Polson S. W., Godiska R., & Mead D. (2010). Functional viral metagenomics and the next generation of molecular tools. *Trends in microbiology*, *18*(1), 20–9.

Schoenfeld T., Patterson M., Richardson P. M., Wommack K. E., Young M., & Mead D. (2008). Assembly of viral metagenomes from yellowstone hot springs. *Applied and environmental microbiology*, *74*(13), 4164–4174.

Seed K. D., Lazinski D. W., Calderwood S. B., & Camilli A. (2013). A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature*, *494*(7438), 489–491.

Sharman M., Thomas J. E., Skabo S., & Holton T. a. (2008). Abacá bunchy top virus, a new member of the genus Babuvirus (family Nanoviridae). *Archives of virology*, *153*(1), 135–47.

Sharon I., Alperovitch A., Rohwer F., Haynes M., Glaser F., Atamna-Ismaeel N., Pinter R. Y., Partensky F., Koonin E. V, Wolf Y. I., Nelson N., & Béjà O. (2009). Photosystem I gene cassettes are present in marine virus genomes. *Nature*, *461*(7261), 258–62.

Sharon I., Battchikova N., Aro E.-M., Giglione C., Meinnel T., Glaser F., Pinter R. Y., Breitbart M., Rohwer F., & Béjà O. (2011). Comparative metagenomics of microbial traits within oceanic viral communities. *The ISME journal*, *5*(7), 1178–90.

Shokralla S., Spall J. L., Gibson J. F., & Hajibabaei M. (2012). Next-generation sequencing technologies for environmental DNA research. *Molecular ecology*, *21*(8), 1794–805.

Short C. M., Rusanova O., & Short S. M. (2011). Quantification of virus genes provides evidence for seed-bank populations of phycodnaviruses in Lake Ontario, Canada. *The ISME journal*, *5*(5), 810–821.

Short C. M., & Suttle C. A. (2005). Nearly identical bacteriophage structural gene sequences are widely distributed in both marine and freshwater environments. *Applied and environmental microbiology*, *71*(1), 480–486.

Short S. M., & Short C. M. (2009). Quantitative PCR reveals transient and persistent algal viruses in Lake Ontario, Canada. *Environmental microbiology*, *11*(10), 2639–48.

Sime-Ngando T., Colombet J., Personnic S., Domaizon I., Dorigo U., Perney P., Hustache J. C., Viollier E., & Jacquet S. (2007). Short-term variations in abundances and potential activities of viruses, bacteria and nanoprotists in Lake Bourget. *Ecological Research*, *23*(5), 851–861.

Sime-Ngando T., Lucas S., Robin A., Tucker K. P., Colombet J., Bettarel Y., Desmond E., Gribaldo S., Forterre P., Breitbart M., & Prangishvili D. (2010). Diversity of virus-host systems in hypersaline Lake Retba, Senegal. *Environmental microbiology*, *13*(8), 1956–1972.

Slimani M., Pagnier I., Raoult D., & La Scola B. (2013). Amoebae as battlefields for bacteria, giant viruses, and virophages. *Journal of virology*, *87*(8), 4783–5.

Smits S. a, & Ouverney C. C. (2010). jsPhyloSVG: a javascript library for visualizing interactive and vector-based phylogenetic trees on the web. *PLoS One*, *5*(8), e12267.

Snyder J. C., Wiedenheft B., Lavin M., Roberto F. F., Spuhler J., Ortmann A. C., Douglas T., & Young M. (2007). Virus movement maintains local virus population diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(48), 19102–7.

Solonenko S. a, Ignacio-Espinoza J. C., Alberti A., Cruaud C., Hallam S., Konstantinidis K., Tyson G., Wincker P., & Sullivan M. B. (2013). Sequencing platform and library preparation choices impact viral metagenomes. *BMC genomics*, *14*, 320.

Sorokin V. a, Gelfand M. S., & Artamonova I. I. (2010). Evolutionary dynamics of clustered irregularly interspaced short palindromic repeat systems in the ocean metagenome. *Applied and environmental microbiology*, *76*(7), 2136–44.

Stern A., & Sorek R. (2011). The phage-host arms race: shaping the evolution of microbes. *BioEssays*, *33*(1), 43–51.

Steward G. F., & Preston C. M. (2011). Analysis of a viral metagenomic library from 200 m depth in Monterey Bay, California constructed by direct shotgun cloning. *Virology journal*, *8*(1), 287.

Sullivan M. B., Coleman M. L., Weigele P., Rohwer F., & Chisholm S. W. (2005). Three Prochlorococcus cyanophage genomes: signature features and ecological interpretations. *PLoS biology*, *3*(5), e144.

Sun S., Chen J., Li W., Altintas I., Lin A., Peltier S., Stocks K., Allen E. E., Ellisman M., Grethe J., & Wooley J. (2011). Community cyberinfrastructure for Advanced Microbial Ecology Research and Analysis: the CAMERA resource. *Nucleic acids research*, *39*(Database issue), D546–51.

Suttle C. A. (1994). The Significance of Viruses to Mortality in Aquatic Microbial Communities. *Microbial ecology*, *28*, 237–243.

Suttle C. A. (2005). Viruses in the sea. *Nature*, *437*(7057), 356–61.

Suttle C. A. (2007). Marine viruses--major players in the global ecosystem. *Nature Reviews Microbiology*, *5*(10), 801–812.

Svraka S., Rosario K., Duizer E., van der Avoort H., Breitbart M., & Koopmans M. (2010). Metagenomic sequencing for virus identification in a public-health setting. *The Journal of general virology*, *91*(Pt 11), 2846–56.

T

Takemura M. (2001). Poxviruses and the origin of the eukaryotic nucleus. *Journal of molecular evolution*, *52*(5), 419–25.

Terns M. P., & Terns R. M. (2011). CRISPR-based adaptive immune systems. *Current opinion in microbiology*, *14*(3), 321–7.

Thingstad T. F. (2000). Elements of a theory for the mechanisms controlling abundance, diversity, and biogeochemical role of lytic bacterial viruses in aquatic systems. *Limnology and Oceanography*, *45*, 1320–1328.

Thomas R., Grimsley N., Escande M.-L., Subirana L., Derelle E., & Moreau H. (2011). Acquisition and maintenance of resistance to viruses in eukaryotic phytoplankton populations. *Environmental microbiology*, *13*(6), 1412–20.

Thompson L. R., Zeng Q., Kelly L., Huang K. H., Singer A. U., Stubbe J., & Chisholm S. W. (2011). Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), E757–64.

Thyrhaug R., Larsen A., Thingstad T. F., & Bratbak G. (2003). Stable coexistence in marine algal host-virus systems. *Marine Ecology Progress Series*, *254*, 27–35.

Tucker K. P., Parsons R., Symonds E. M., & Breitbart M. (2011). Diversity and distribution of single-stranded DNA phages in the North Atlantic Ocean. *The ISME journal*, *5*(5), 822–30.

V

Van Regenmortel M. H. V. (2003). Viruses are real, virus species are man-made, taxonomic constructions. *Archives of virology*, *148*(12), 2481–8.

Veesler D., Quispe J., Grigorieff N., Potter C. S., Carragher B., & Johnson J. E. (2012). Maturation in action: CryoEM study of a viral capsid caught during expansion. *Structure*, *20*(8), 1384–90.

Vega Thurber R. (2009). Current insights into phage biodiversity and biogeography. *Current opinion in microbiology*, *12*(5), 582–587.

Vega Thurber R., Haynes M., Breitbart M., Wegley L., & Rohwer F. (2009a). Laboratory procedures to generate viral metagenomes. *Nature protocols*, *4*(4), 470–483.

Vega Thurber R. L., Barott K. L., Hall D., Liu H., Rodriguez-Mueller B., Desnues C., Edwards R. A., Haynes M., Angly F. E., Wegley L., & Rohwer F. L. (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(47), 18413–8.

Vega Thurber R., Willner-Hall D., Rodriguez-Mueller B., Desnues C., Edwards R. A., Angly F., Dinsdale E., Kelly L., & Rohwer F. (2009b). Metagenomic analysis of stressed coral holobionts. *Environmental microbiology*, *11*(8), 2148–63.

Victoria J. G., Kapoor A., Li L., Blinkova O., Slikas B., Wang C., Naeem A., Zaidi S., & Delwart E. (2009). Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *Journal of virology*, *83*(9), 4642–51.

Villarreal L. P. (2004). Are viruses alive? *Scientific American*, *291*(6), 100–5.

Villarreal L. P., & Witzany G. (2010). Viruses are essential agents within the roots and stem of the tree of life. *Journal of theoretical biology*, *262*(4), 698–710.

W

Wang J., Yang D., Zhang Y., Shen J., van der Gast C., Hahn M. W., & Wu Q. (2011). Do patterns of bacterial diversity along salinity gradients differ from those observed for macroorganisms? *PLoS One*, *6*(11), e27597.

Weinbauer M. G., Brettar I., & Ho M. G. (2003). Lysogeny and virus-induced mortality of bacterioplankton in surface, deep, and anoxic marine waters. *Limnology and Oceanography*, *48*(4), 1457–1465.

Weitz J. S., Poisot T., Meyer J. R., Flores C. O., Valverde S., Sullivan M. B., & Hochberg M. E. (2012). Phage-bacteria infection networks. *Trends in microbiology*, *21*(2).

Whitaker R. J. (2006). Allopatric origins of microbial species. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *361*(1475), 1975–84.

Whon T. W., Kim M.-S., Roh S. W., Shin N.-R., Lee H.-W., & Bae J.-W. (2012). Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *Journal of virology*, *86*(15), 8221–8331.

Wilber A. W., Doye J. P. K., Louis A. A., & Lewis A. C. F. (2009). Monodisperse self-assembly in a model with protein-like interactions. *Journal of Chemical Physics*, *131*(175102), 1–11.

Wilhelm S. W., Carberry M. J., Eldridge M. L., Poorvin L., Saxton M. a, & Doblin M. A. (2006). Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of g20 genes. *Applied and environmental microbiology*, *72*(7), 4957–4963.

Williams W. D. (1998). Salinity as a determinant of the structure of biological communities in salt lakes. *Hydrobiologia*, *180269*(180269), 191–201.

Williamson S. J., Allen L. Z., Lorenzi H. a, Fadrosh D. W., Brami D., Thiagarajan M., McCrow J. P., Tovchigrechko A., Yooseph S., & Venter J. C. (2012). Metagenomic Exploration of Viruses throughout the Indian Ocean. *PLoS One*, *7*(10), e42047.

Williamson S. J., Cary S. C., Williamson K. E., Helton R. R., Bench S. R., Winget D., & Wommack K. E. (2008a). Lysogenic virus-host interactions predominate at deep-sea diffuse-flow hydrothermal vents. *The ISME journal*, *2*(11), 1112–21.

Williamson S. J., Rusch D. B., Yooseph S., Halpern A. L., Heidelberg K. B., Glass J. I., Andrews-Pfannkoch C., Fadrosh D., Miller C. S., Sutton G., Frazier M., & Venter J. C. (2008b). The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS One*, *3*(1), e1456.

Willner D., Furlan M., Haynes M., Schmieder R., Angly F. E., Silva J., Tammadoni S., Nosrat B., Conrad D., & Rohwer F. (2009a). Metagenomic analysis of respiratory tract DNA viral communities in cystic fibrosis and non-cystic fibrosis individuals. *PLoS One*, *4*(10), e7370.

Willner D., & Hugenholtz P. (2013). From deep sequencing to viral tagging: Recent advances in viral metagenomics. *BioEssays*, 1–7.

Willner D., Thurber R. V., & Rohwer F. (2009b). Metagenomic signatures of 86 microbial and viral metagenomes. *Environmental microbiology*, *11*(7), 1752–1756.

Winter C., Matthews B., & Suttle C. A. (2013). Effects of environmental variation and spatial distance on Bacteria, Archaea and viruses in sub-polar and arctic waters. *The ISME journal*, 1–12.

Wommack K. E., Bhavsar J., Polson S. W., Chen J., Dumas M., Srinivasiah S., Furman M., Jamindar S., & Nasko D. J. (2012). VIROME: a standard operating procedure for analysis of viral metagenome sequences. *Standards in Genomic Sciences*, *6*(3), 427–439.

Wommack K. E., Bhavsar J., & Ravel J. (2008). Metagenomics: read length matters. *Applied and environmental microbiology*, *74*(5), 1453–1463.

Wylie K. M., Weinstock G. M., & Storch G. a. (2012). Emerging view of the human virome. *Translational research : the journal of laboratory and clinical medicine*, *160*(4), 283–90.

Y

Yin Y., & Fischer D. (2008). Identification and investigation of ORFans in the viral world. *BMC genomics*, *9*, 24.

Yip K. Y., Cheng C., & Gerstein M. (2013). Machine learning and genome annotation: a match meant to be? *Genome biology*, *14*(5), 205.

Yoon H. S., Hackett J. D., Van Dolah F. M., Nosenko T., Lidie K. L., & Bhattacharya D. (2005). Tertiary endosymbiosis driven genome evolution in dinoflagellate algae. *Molecular biology and evolution*, *22*(5), 1299–308.

Yoon H. S., Price D. C., Stepanauskas R., Rajah V. D., Sieracki M. E., Wilson W. H., Yang E. C., Duffy S., & Bhattacharya D. (2011). Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science*, *332*(6030), 714–7.

Yooseph S., Sutton G., Rusch D. B., Halpern A. L., Williamson S. J., Remington K., Eisen J. a, Heidelberg K. B., Manning G., Li W., Jaroszewski L., Cieplak P., Miller C. S., Li H., Mashiyama S. T., Joachimiak M. P., van Belle C., Chandonia J.-M., Soergel D. a, Zhai Y., Natarajan K., Lee S., Raphael B. J., Bafna V., Friedman R., Brenner S. E., Godzik A., Eisenberg D., Dixon J. E., Taylor S. S., Strausberg R. L., Frazier M., & Venter J. C.

(2007). The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS biology*, *5*(3), e16.

Yoshida M., Takaki Y., Eitoku M., Nunoura T., & Takai K. (2013). Metagenomic Analysis of Viral Communities in (Hado) Pelagic Sediments. *PLoS One*, *8*(2), e57271.

Z

Zaitlin M. (1998). The Discovery of the Causal Agent of the Tobacco Mosaic Disease. In S. Kung & S. Yang (Eds.), *Discoveries in Plant Biology* (World Publ., pp. 105–110). Hong Kong.

Zhao Y., Temperton B., Thrash J. C., Schwalbach M. S., Vergin K. L., Landry Z. C., Ellisman M., Deerinck T., Sullivan M. B., & Giovannoni S. J. (2013). Abundant SAR11 viruses in the ocean. *Nature*.

Curriculum vitae scientifique

Publications Scientifiques

Publications en préparation

Wood-Charlson E. M., Weynberg K. D., Suttle C. A., **Roux S.**, van Oppen M.J.H. Methodological biases in coral viromics. Soumis à *Plos ONE*.

Roux S., Tournayre J., Mahul A., Debroas D. and Enault F. Metavir 2: Comparison of viral metagenomes and analysis of assembled datasets. Soumis à *BMC Bioinformatics.*

Roux S., Enault F., Bronner G., Vaulot D., Forterre P. and Krupovic M. Chimeric viruses blur the border between the major groups of eukaryotic single-stranded DNA viruses. En révision, *Nature Communications*.

Roux S., Krupovic M., Debroas D., Forterre P. and Enault F. Uncontaminated viromes reveal the abundance and diversity of metabolism genes in environmental viruses. Soumis à *The ISME Journal*.

Publications parues ou sous presse dans des revues à comité de lecture

Roux S., Enault F., Debroas D. Application des approches métagénomiques à l'étude de la diversité virale environnementale. *Virologie* (Article en français). 2013

Roux S., Krupovic M, Poulet A, Debroas D, Enault F (2012) Evolution and Diversity of the Microviridae Viral Family through a Collection of 81 New Complete Genomes Assembled from Virome Reads. *PLoS ONE* 7(7): e40418. doi:10.1371/journal.pone.0040418. 2012

Roux S., Enault F, Robin A, Ravet V, Personnic S, Theil S, Colombet J, Sime-Ngando T, Debroas D. Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS ONE* 7(3): e33641. doi:10.1371/journal.pone.0033641. 2012

Roux S., Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D, Enault F. Metavir: a web server dedicated to virome analysis. *Bioinformatics.* 27(21):3074-5. 2011

Roux S., Enault F., Bronner G., Debroas D. Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (Bacteria and Archaea) in

ecosystems. *FEMS Microbiol Ecol.* 78(3):617-28. doi: 10.1111/j.1574-6941.2011.01190.x. 2011

Communications à des congrès nationaux et internationaux

Roux S., Debroas D., Mahul A., Enault F. Metavir, a web server dedicated to virome analysis : presentation and hands-on training. *Environmental Virology Workshop*. Tucson, Arizona

Roux S., Enault F., Ravet V., Mahul A., Sime-Ngando T., Debroas D. Assessment of viral communities richness, diversity, and biogeography through viromes comparative analyses. *Viruses of microbes*. Bruxelles, Belgique. (Communication affichée)

Roux S., Krupovic M., Poulet A., Debroas D., Enault F. Redéfinition d'une famille virale par l'assemblage de génomes complets à partir de données métagénomiques. *Journées de l'école doctorale Sciences de la Vie, Santé, Agronomie et Environnement*. Clermont-Ferrand

Roux S., Enault F., Ravet V., Mahul A., Sime-Ngando T., Debroas D. Étude des communautés virales par approche métagénomique. *Rencontre des microbiologiste clermontois*. Clermont-Ferrand

Roux S., Taib N., Mangot J.F., Hugoni M., Mary I., Ravet V., Bronner G., Enault F., Debroas D. Analyse des données de séquençage massif par des méthodes phylogénétiques : outils bioinformatiques dédiés et applications. *Colloque Génomique Environnementale*, Lyon

Roux S., Enault F., Robin A., Ravet V., Personnic S., Theil S., Colombet J., Sime-Ngando T., Debroas D. Metagenomic analysis of the viral communities from temperate freshwater lakes. *Viruses of the environment*. Heidelberg, Germany

Roux S. Analysis of metagenomes from lacustrin viral communities. *ALPHY : Evolution genomic, Bioinformatics, Alignment and Phylogenies*. Lyon, France.

Activités complémentaires

2010 – 2013 : Enseignement, service de monitorat (64h/an) effectué dans le cadre de modules d'Initiation à la Bioinformatique, Statistiques, Evolution et Programmation.

2013 : Reviewer pour les revues *Applied and Environmental Microbiology* et *Plos ONE*

Annexes

Annexe A.1 : Article

Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (*Bacteria* and *Archaea*) in ecosystems

Simon Roux[1,2], François Enault[1,2], Gisele Bronner[1,2] and Didier Debroas[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand

[2]CNRS, UMR 6023, LMGE, F-63177 Aubière

[3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France

Publié en décembre 2011 dans **FEMS Microbiology Ecology** (78, 3 : 617-628)

RESEARCH ARTICLE

Comparison of 16S rRNA and protein-coding genes as molecular markers for assessing microbial diversity (*Bacteria* and *Archaea*) in ecosystems

Simon Roux[1,2], François Enault[1,2], Gisèle Bronner[1,2] & Didier Debroas[1,2]

[1]Clermont Université, Université Blaise Pascal, Laboratoire 'Microorganismes: Génome et Environnement', Clermont-Ferrand, France; and [2]CNRS, UMR 6023, LMGE, Aubiere, France

Correspondence: Didier Debroas, Universite Blaise Pascal, Laboratoire de Biologie des Protistes – UMR CNRS 6023, Aubiere 63177, France. Tel.: +334 7340 7837; fax: +334 7340 7670; e-mail: didier.debroas@univ-bpclermont.fr

Received 19 June 2011; revised 12 August 2011; accepted 17 August 2011.
Final version published online 19 September 2011.

DOI: 10.1111/j.1574-6941.2011.01190.x

Editor: Tillmann Luders

Keywords
diversity; 16S rRNA gene; protein-coding genes; aquatic ecosystems.

Abstract

PCR amplification of the rRNA gene is the most popular method for assessing microbial diversity. However, this molecular marker is often present in multiple copies in cells presenting, in addition, an intragenomic heterogeneity. In this context, housekeeping genes may be used as taxonomic markers for ecological studies. However, the efficiency of these protein-coding genes compared to 16S rRNA genes has not been tested on environmental data. For this purpose, five protein marker genes for which primer sets are available, were selected (rplB, pyrG, fusA, leuS and rpoB) and compared with 16S rRNA gene results from PCR amplification or metagenomic data from aquatic ecosystems. Analysis of the major groups found in these ecosystems, such as *Actinobacteria*, *Bacteroides*, *Proteobacteria* and *Cyanobacteria*, showed good agreement between the protein markers and the results given by 16S rRNA genes from metagenomic reads. However, with the markers it was possible to detect minor groups among the microbial assemblages, providing more details compared to 16S rRNA results from PCR amplification. In addition, the use of a set of protein markers made it possible to deduce a mean copy number of rRNA operons. This average estimate is essentially lower than the one estimated in sequenced genomes.

Introduction

The functional and species diversity of microorganisms has been shaped by 3.5 billion years of evolution, enabling them to colonize all aquatic ecosystems, even the most extreme (Thornburg *et al.*, 2010). Microorganisms are involved in all the basic processes, from degrading organic matter to regulating the composition of the Earth's atmosphere equilibria such as O_2–CO_2 or CH_4. Despite playing this crucial functional role in the terrestrial ecosystem, there has been only limited progress in identifying and classifying prokaryotes, as only 1% of microbes can be cultivated with classical microbial methods (Amann *et al.*, 1995). Over the past two decades, the use of techniques based on ribosomal RNA (rRNA) has revolutionized knowledge on the microorganisms present in ecosystems. Microbial diversity studies are now dominated by approaches involving techniques such as cloning-sequencing, fluorescent *in situ* hybridization and genomic fingerprinting (e.g. DGGE: Denaturing Gradient Gel Electrophoresis, T-RFLP), revealing the broad diversity of microbial communities. Phylogenetic reconstruction based on rRNA has made it possible to highlight the existence of new clades specific to certain ecosystems, such as a highly abundant clade known as SAR11 that is found in all the oceans (Morris *et al.*, 2002). Similarly, *Archaea* have been identified in the euphotic zones of marine ecosystems (DeLong, 1998). Finally, SSU rRNA sequences have made it possible to define 40 phyla, for half of which no bacteria have been isolated or cultivated (Hugenholtz, 2002).

Despite these advances in assessing microbial diversity, 16S rRNA gene-based approaches are highly questionable. PCR amplification of 16S rRNA gene sequences from samples has been shown to miss half of rRNA bacterial diversity (Hong *et al.*, 2009). Metagenomics is not subject

FEMS MICROBIOLOGY ECOLOGY

to this amplification bias, but 16S rRNA gene sequences represent only a small part of these datasets. Furthermore, the 16S rRNA gene is not the only phylogenetic marker available, and possibly not the best one. Recent works have highlighted housekeeping genes, coding for ribosomal proteins, DNA-linked protein or all the amino-acyl synthetases, as robust phylogenetic markers (Santos & Ochman, 2004; von Mering *et al.*, 2007; Case *et al.*, 2007; Konstantinidis *et al.*, 2006; Wu *et al.*, 2011). Furthermore, in a study of 111 completely sequenced bacterial genomes, Case *et al.* (2007) showed that the rpoB gene provided more phylogenetic resolution than the 16S rRNA gene. Using a pyrosequencing approach, Schellenberg *et al.* (2009) highlighted that chaperonin-60 amplicons improved species resolution over the 16S rRNA target. These phylogenetic markers are present in a single copy, unlike the 16S rRNA gene, which is known to be present in multiple copies (Coenye & Vandamme, 2003), thus creating a bias for diversity studies. In addition, 16S rRNA gene copies can be heterogeneous and a single species could produce complex DGGE patterns, similar to those obtained with an environmental assemblage (Dahllöf *et al.*, 2000). Therefore, as underlined for the rpoB gene, housekeeping genes may be used as taxonomic markers for ecological studies due to (1) the presence of slow and fast-evolving regions, (2) a low rate of lateral gene transfer, and (3) a single copy per genome. Thus, using multiple conserved protein-coding regions in association with SSU rRNA certainly appears to be the best way to assess ecosystem diversity. However, these markers have often been tested on cultivated bacteria (Case *et al.*, 2007) and have not yet been proven to be suitable for microbial ecology, whereas numerous published metagenomic studies could offer an interesting dataset for this purpose.

Here, the phylogenetic markers proposed by Santos & Ochman (2004) and for which primers were designed, were used to study bacterial and archaeal diversities in metagenomic studies. Six metagenomes were selected from the GOS project data obtained in contrasted environments (coastal, estuary, ocean and lake) for which 16S rRNA clone libraries were available (Shaw *et al.*, 2008), enabling us to compare diversity as deduced from selected phylogenetic markers present in the metagenomic library against the diversity obtained on the 16S rRNA amplicons.

Materials and methods

Alignment and phylogenetic analysis of reference sequences for protein markers

Among the 10 candidate phylogenetic markers for which primers were available (Santos & Ochman, 2004), different housekeeping functions were tested to check that all these type of genes could be used as phylogenetic markers. The final set was composed of a ribosomal protein (ribosomal protein L2, rplB), a protein implicated in nucleotide metabolism (Cytidine triphosphate synthase, pyrG), a translation protein (translation elongation factor G, fusA), a protein associated with a tRNA (leucyl-tRNA synthetase, leuS), and a transcription protein (RNA polymerase Beta, rpoB) (Supporting Information, Table S1).

Protein sequences were extracted from the KEGG database (Kanehisa, 2002) and reduced to one sequence per bacterial genus. To perform accurate phylogenetic affiliations for the most-retrieved groups, specific alignments reduced to *Alphaproteobacteria*, *Betaproteobacteria* and *Actinobacteria*, were generated for each marker. These alignments included all the sequences available in KEGG for these classes. Reference alignments were performed with MUSCLE (Edgar, 2004), with poorly aligned positions and divergent regions further excluded using GBLOCKS (Castresana, 2000). Over-diverging sequences were removed manually. Phylogenetic trees were then generated with PHYML (Guindon & Gascuel, 2003), using the Jones-Taylor-Thornton substitution model (Jones *et al.*, 1992) with automatic evaluation of gamma parameter and proportion of invariable sites. Monophyly of the main taxonomic phyla was checked manually.

As reference sequence coverage by metagenomic fragments is often incomplete, the uniformity of the phylogenetic information available throughout the sequence was evaluated to guarantee the homogeneity of the observed signal, whatever the region covered by the metagenomic fragment. Alignments were consequently screened with PAML (Yang, 2007) to estimate substitution rate constancy over a sliding-window of 20 positions [rates were defined as categories from 1 (low) to 10 (high)]. Parts of alignments presenting low or high substitution rates, indicating a potentially insufficient or saturated phylogenetic signal, were removed. FusA alignment showed a low evolutionary rate segment (from position 0–800) matching a guanosine triphosphate-binding domain, which was therefore excluded. As the other four alignments presented no such parts, all positions were retained.

Metagenome search for protein markers

Six metagenomes were selected from the GOS project data obtained in contrasted environments (Rusch *et al.*, 2007) and for which 16S rRNA clone libraries were available (Shaw *et al.*, 2008): GS08 (Newport Harbor; coastal), GS11 (Delaware Bay; estuary), GS12 (Chesapeake Bay; estuary), GS19 (Caribbean Sea; coastal), GS20 (Gatun Lake; freshwater) and GS22 (Pacific Ocean; open ocean).

An automatic pipeline was developed to generate a taxonomic affiliation from metagenomic reads (Fig. S1) and

is available on demand. Briefly, complete metagenomes were first compared with a reduced database formatted from the whole protein sequences of the marker genes taken from their associated domain in the Protein Family database (Bateman *et al.*, 1999). Then, a second BLASTX was performed against the NCBI no. database for fragments with an e-value lower than 0.001 in the first BLAST. Fragments were only considered homologous to the screened marker if their most similar hit in this second BLAST was a screened marker sequence with a local alignment greater than 100 residues and an e-value lower than 0.001. This local alignment was extended to other local similarities detected on the same frame, and the resulting DNA sequence was retrieved and translated in protein sequence. Best BLAST hit taxonomic affiliations were taken from this second BLAST step.

Taxonomic affiliation of metagenomic fragments

Metagenomic sequences were included within the reference alignment of each marker using HMMER (Eddy, 1998), and GBLOCKS was then used to restrict alignments to informative positions (minimal block size of five positions, no gap allowed). Cleaned alignments were screened for a minimum length evaluated for each marker from simulation results and phylogenetically analyzed. Trees were computed with PHYML using the JTT model and automatic evaluation of gamma parameter and proportion of invariable sites. The set of trees (one for each homologous fragment) was handled with a custom-designed Java script in order to retrieve the phylogenetic group the fragment was inserted in for each tree. The metagenomic fragment is then affiliated to the phylum of the group of reference sequences that share the last common ancestor with it, if this group is taxonomically consistent. If not, the fragment was considered 'unclassified'. One hundred bootstrap trees were computed for each detection considered 'ambiguous' (i.e. phylum detected by protein marker but not by the 16S rRNA gene, or detected by fewer than three markers including the 16S rRNA gene) using PHYML with the previously described parameter.

This procedure was tested with metagenomic fragments simulated from protein sequences from KEGG that were not included in reference alignments. For each protein marker, 500 sequences plus a start position and fragment size (from 100 to 330 amino acids) were randomly chosen to produce a simulated metagenome. Random fragments were affiliated at the phylum level using the phylogenetic approach described above and these affiliations were compared to best BLAST hit affiliations (Table S2). This procedure was then applied to the six metagenomes selected from the GOS project (Rusch *et al.*, 2007).

16S rRNA affiliations and operon copy number estimation

16S rRNA genes were detected in the metagenomes using RNA_HMM3 (Huang *et al.*, 2009). The sequences were aligned with the SINA WEB aligner from SILVA database (http://www.arb-silva.de/) (Pruesse *et al.*, 2007) and phylogenetically analyzed using the software ARB (Ludwig *et al.*, 2004) with a SILVA database.

The rRNA operon copy number was estimated for each phylum as the ratio between the number of sequences of protein markers detected in metagenomes and the nucleotidic size of this marker, divided by the same ratio for 16S rRNA gene sequences detected in metagenomes. As low detection levels are likely to introduce some bias in these calculated numbers, only groups detected more than twice by at least three protein markers were considered for each metagenome. As protein markers are known to be present in a single copy per genome, these ratios are considered to be reliable estimators of the mean copy number of rRNA operon.

Results

Richness and diversity inferred by protein markers and 16S rRNA genes

Six GOS Project metagenomes from various aquatic environments for which 16S PCR amplifications had been conducted were analyzed using both phylogenetic and BLAST approaches (Fig. 1, Table 1, Fig. S2). The average number of affiliations ranged from 72 to 592 (average: 200) per marker per metagenome. In general, BLAST and phylogenetic tree affiliations were congruent (results not shown); however, the phylogenetic method recovered more affiliations than BLAST did. Thus, some groups, such as *Verrucomicrobia* in G20, were detected by tree affiliation in GS20, but remained undetected for all protein markers using the best BLAST hit.

A presence/absence analysis of each marker was conducted for each metagenome. This allowed a global comparison of the taxonomic diversities estimated by affiliation from protein markers, and from 16S rRNA detected in metagenomes or from PCR-amplified 16S-rRNA (Table 1). In this last case, only bacterial rRNA genes were amplified (Shaw *et al.*, 2008).

In Table 1, two categories of phyla can be distinguished. The first category corresponds to taxonomic groups which were detected by all phylogenetic markers, i.e. all five protein markers as well as the 16S rRNA gene

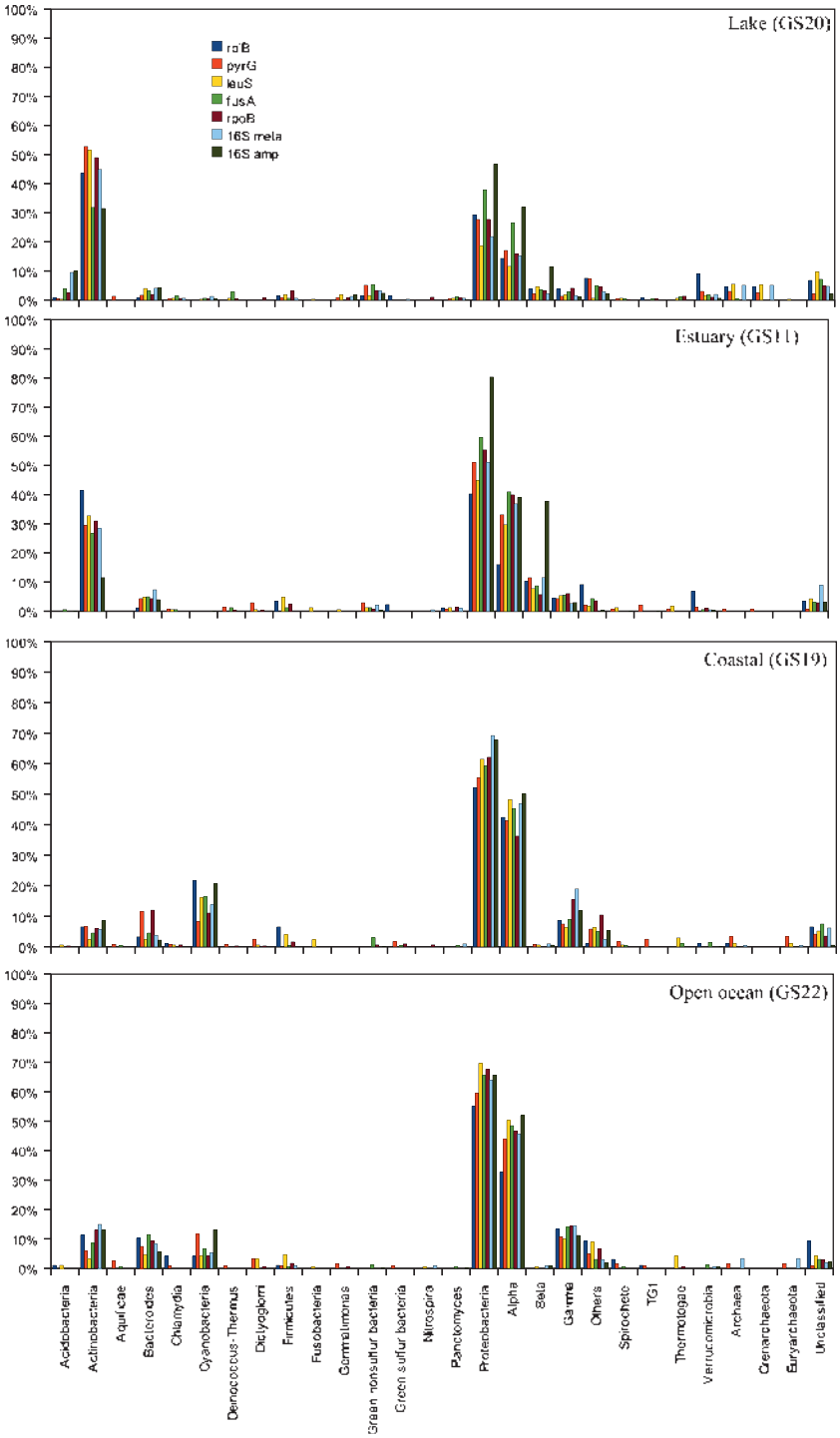


Fig. 1. Taxonomic affiliation for four contrasted aquatic ecosystems, with five protein markers (rplB, pyrG, leuS, fusA and rpoB), metagenomic 16S rRNA and PCR-amplified 16S rRNA genes.

in the metagenomes (marked by #) and clone library post-amplification (marked by *). For example, *Actinobacteria*, *Bacteroides* and *Proteobacteria* belonged to this category in all environmental samples. The second category corresponded to bacterial or archaeal groups that were only detected by some markers. Thus, some groups were detected by at least three protein markers and were not associated with 16S rRNA gene detection, such as *Dictyoglomi* (GS11, 19 and 22) and *Deinococcus thermus* (GS11, 12 and 20). Other groups were PCR-amplified, whereas no 16S rRNA gene was detected in the metagenome studied, such as *Acidobacteria* (GS11) or *Planctomyces* (GS8,

Table 1. Taxonomic affiliation computed at a phylum level for 16S rRNA gene and the five protein markers

	Lake	Estuary		Coastal		Open ocean
	GS20	GS11	GS12	GS08	GS19	GS22
Acidobacteria	#5*	1*	1	1	2	2
Actinobacteria	#5*	#5*	#5*	#5*	#5*	#5*
Aquificae	1	0	0	0	2	2
Bacteroides	#5*	#5*	#5*	#5*	#5*	#5*
Chlamydia	#4	3	1	1	4	2
Cyanobacteria	#3*	0*	0	#2*	#5*	#5*
Deinococcus-Thermus	3	3	3	2	2	1
Dictyoglomi	1	3	1	1	3	3
Firmicutes	#5	4	4	#4	4	#5
Fusobacteria	1	1	1	2	1	1
Gemmatimonas	#3*	1	1*	#0	0	2
Green nonsulfur bacteria	#5*	#4*	#4*	#3*	2*	1*
Green sulfur bacteria	#1	1	0	2	3	1
Nitrospira	1	#0	0	#0	1	#1
Planctomyces	#4*	#4	1*	3*	#1	1*
Proteobacteria	#5*	#5*	#5*	#5*	#5*	#5*
Alpha	#5*	#5*	#5*	#5*	#5*	#5*
Beta	#5*	#5*	#5*	#5*	#2*	#1*
Gamma	#5*	#5*	#5*	#5*	#5*	#5*
Others	#5*	5*	#5*	#5*	#5*	#5*
Spirochete	4	2*	2	1	3	3
TG1	3	1*	1*	2	1	2
Thermotogae	3	2	2	3	2	2
Verrucomicrobia	#5*	#4*	#4*	#4*	2*	#1*
Archaea	#4	1	1	#4	#3	#1
Crenarchaeota	#3	1	1	#3	0	0
Euryarchaeota	1	0	0	#1	#2	#1
Unclassified	#5*	#5*	#5*	#5*	#5*	#5*

A group is considered retrieved when more than five affiliations are made. Number of protein markers which retrieve each group are noted, alongside metagenomic 16S rRNA gene detection (#) and PCR-amplified 16S rRNA gene detection (*).

12 and 22). Conversely, *Firmicutes* were detected in all aquatic environments with at least four markers (six in GS20 and GS22 metagenomes), but there was no 16S rRNA gene sequence associated with this phylogenetic group in the clone libraries obtained by PCR.

The taxonomic groups belonging to the first category often represented the most commonly retrieved groups in aquatic ecosystems, as seen in Fig. 1 (data for GS12 and GS08 and some minor groups can be consulted in Figs S2 and S3). In the estuary (GS11), *Alphaproteobacteria* and *Actinobacteria* were the main phyla, representing 32.0% (±10.0%) and 32.2% (±5.6%) respectively. The coastal ecosystem (GS19) was dominated by *Alphaproteobacteria* and *Cyanobacteria*. Open ocean (GS22) was similarly dominated by *Alphaproteobacteria* for all protein markers (44.4 ± 7.0%). In this ecosystem, *Actinobacteria, Bacteroides* and *Cyanobacteria* were detected as minor groups. Finally, the freshwater sample (GS20) was the only metagenome where *Actinobacteria* was the main phylum (around 45.7 ± 8.5%). Among the *Proteobacteria*, *Alpha-*, *Beta-* and *Gammaproteobacteria* were detected by all markers at lower abundances (17.1%, 3.5% and 2.8%, respectively). *Verrucomicrobia*, *Bacteroidetes* and *Archaea* were also detected, with abundances ranging from 2% to 5%. The *Archaea* kingdom mainly consisted of *Crenarchaeota* (2.4 ± 2.4%). Finally, all protein markers and 16S rRNA genes presented a similar ratio of unclassified sequences (5–10%).

Differences were found between the relative abundance of phyla determined by 16S rRNA gene amplification and the same clade determined by ribosomal and protein markers. Thus, *Proteobacteria* seemed preferentially amplified in lake (GS20) and estuaries (GS11 and 12), whereas *Actinobacteria* were underestimated in the estuary clone library (GS11 and 12).

Community composition at a finer phylogenetic level

Because of their high abundance in aquatic ecosystems, therefore representing a reasonable amount of sequences in the metagenomes, and to investigate the causes of discrepancies in abundance estimation from amplified 16S rRNA gene vs. metagenomic fragment affiliations, *Alpha-* and *Betaproteobacteria* as well as *Actinobacteria* classes were analyzed further using specific datasets of these groups.

Alphaproteobacteria

Alphaproteobacteria was the main group detected in GS08, GS11, GS12, GS19 and GS22. The *Alphaproteobacteria* communities were dominated by the order of *Rickettsiales*, notably members of the *Pelagibacter* genus, with the exception of GS08 (Table 2). For GS11, GS12, GS19, GS20 and GS22, members of the *Pelagibacter* genus account for over 75% of total *Alphaproteobacteria* for all markers, even rising to over 90% (especially in GS12). Thus, *Alphaproteobacteria* communities could prove highly uniform in aquatic environments, regardless of the type of ecosystem studied (coastal, estuary, freshwater, open ocean). The GS08 sample appeared to represent a separate subgroup: using protein markers, most of *Alphaproteobacteria* were either affiliated to *Pelagibacter* or were 'unclassified' (Table 2, Table S3), whereas with the 16S rRNA gene, members of *Rhodobacter* genus were retrieved, which is notably the main group detected with PCR-amplified sequences. Hence, there were important differences in *Alphaproteobacteria* affiliations between 16S rRNA gene and protein markers for this specific sample.

Table 2. Results of affiliation to genus *Pelagibacter* for sequences initially associated with *Alphaproteobacteria*

	Lake	Estuary		Coastal		Open ocean
	GS20 (%)	GS11 (%)	GS12 (%)	GS08 (%)	GS19 (%)	GS22 (%)
rplB	100.0	78.6	92.0	40.0	84.6	90.6
pyrG	75.0	89.1	94.2	14.3	86.0	84.9
leuS	86.8	87.8	98.0	30.8	77.4	89.6
fusA	49.3	83.3	75.8	8.3	71.4	73.8
rpoB	80.5	86.5	87.7	60.0	82.6	81.1
Metagenomic 16S rRNA	81.3	95.7	92.7	67.7	71.3	77.2
PCR-amplified 16S rRNA	93.9	92.6	95.6	17.3	79.4	72.3

Betaproteobacteria

Betaproteobacteria was one of the main classes retrieved for three metagenomes (GS11, GS12 and GS20, i.e. the two estuary samples and the freshwater sample). Generally, the *Betaproteobacteria* communities were similar in these three ecosystems and were mainly composed of members of *Burkholderia* and unclassified *Betaproteobacteria* (regardless of affiliation method or marker used; Table 3). Note that the *Methylophylales* order has been detected in all three metagenomes with rpoB, metagenomic 16S rRNA gene and PCR-amplified 16S rRNA gene. This order remained undetected with rplB, pyrG and fusA, and was only retrieved in GS11 for leuS. Thus, bacteria related to *Methylophylales* are likely to be present in the three samples, but the absence of affiliation with several markers would indicate a distant relationship between environmental and known members of the *Methylophylales* order.

Actinobacteria

Actinobacteria was main phylum retrieved for the freshwater metagenome (GS20), and was consistently retrieved in all other samples. However, new affiliations of sequences associated to *Actinobacteria* could not provide useful information, as the vast majority were unaffiliated, i.e. emerging at the root of the *Actinobacteria* tree, whatever the marker or metagenome. This high number of unaffiliated sequences may reflect a major divergence between *Actinobacteria* strains entirely sequenced from the public database and the current members of the *Actinobacteria* phylum from the aquatic environment.

Determination of rRNA operon copy number

Per cell rRNA operon copy numbers were estimated from taxonomic groups for which at least two affiliations were retrieved for three or more protein markers (Table 4).

Table 3. Results of affiliation at the order level for sequences associated with *Betaproteobacteria*

	rplB (%)	pyrG (%)	leuS (%)	fusA (%)	rpoB (%)	Metagenomic 16S rRNA (%)	PCR-amplified 16S rRNA (%)
GS11							
Burkholderia	66.7	75.0	62.5	50.0	50.0	54.5	59.5
Unclassified	33.3	25.0	25.0	35.7	41.7	13.6	5.2
Methylophilales	–	–	12.5	–	8.3	31.8	35.3
Other groups	–	–	–	14.3	–	–	–
GS12							
Burkholderia	50.0	25.0	20.0	45.5	27.3	56.0	72.2
Unclassified	50.0	75.0	80.0	54.5	63.6	4.0	3.1
Methylophilales	–	–	–	–	9.1	40.0	24.4
Other groups	–	–	–	–	–	–	0.3
GS20							
Burkholderia	60.0	40.0	16.7	50.0	60.0	57.1	42.1
Unclassified	20.0	60.0	83.3	40.0	20.0	14.3	14.9
Methylophilales	–	–	–	–	20.0	28.6	42.1
Other groups	20.0	–	–	10.0	–	–	0.9

'–' indicates an absence of affiliation for the group with this marker.

Some minor groups with low read number might induce some bias in this estimation, such as *Firmicutes* in lake samples, as the resulting copy number was slightly below 1. An underestimation of the number of 16S rRNA genes can also lead to a copy number lower than 1, e.g. for *Verrucomicrobia* in GS11 and GS08. Generally, the rRNA operon copy number per bacteria was higher in the genomes collected in the rrnDB than in those determined in aquatic ecosystems, with the exception of *Green nonsulfur bacteria*, where this number was equivalent, and *Crenarcheota*. In addition, there were some differences among aquatic ecosystems. For instance, the *Actinobacteria* rRNA copy number was higher in coastal and oceanic samples than in those from estuary or lake. Similarly, a high copy number was detected in GS08 for *Bacteroides*, *Alphaproteobacteria*, *Betaproteobacteria* and *Gammaproteobacteria*. These results have to be linked with the specificity of the *Alphaproteobacteria* community in this sample detected by phylogenetic affiliation. Finally, the copy number for *Crenarchaeota* in the lake sample was higher than those extracted from the genomes collected in the rrnDB, and equivalent to the global 16S rRNA gene copy number detected in *Archaea* genomes.

Discussion

Environmental microorganisms mediate many natural cycles and play a leading role in many ecosystems, particularly in marine environments (Giovannoni & Stingl, 2005). This makes microbial diversity in these environments a major research challenge, both to understand the functioning of these ecosystems and to gain an accurate view of the actors involved. The approach commonly used for diversity estimations involves PCR amplification of the 16S rRNA gene. This experimental strategy gives significantly different estimates of microbial composition when compared to estimations obtained by both 16S rRNA gene or protein-coding sequences of housekeeping genes (known as phylogenetic markers) detected in metagenomes (Venter *et al.*, 2004). Thus, the estimation of species diversity might be strengthened when using additional protein markers for assessing the diversity in metagenomic studies or after amplifying protein coding genes. Moreover, analyzing metagenome-wide 16S rRNA genes neglects a huge amount of information, as very few metagenomic fragments contain 16S rRNA gene sequences (Biers *et al.*, 2009).

Phylogenetic affiliation of protein-coding marker genes

Taxonomic affiliations from metagenomic data are generally based on the best hit obtained using the popular BLAST tool, which provides a fast analysis of many sequences (e.g. Debroas *et al.*, 2009). Nevertheless, Koski & Golding (2001) showed that the closest BLAST hit is not always the nearest phylogenetic neighbor, especially when no close relatives are available in the database. As many bacterial species remain uncultivated, and are therefore missing from the databases, assigning metagenomic sequences using the most similar BLAST hit could lead to diversity assessment bias. Some methods, such as MEGAN (Huson *et al.*, 2009), check the consistency of the taxonomy for a given number of best BLAST hits. However, when few reference sequences are available, the affiliation is likely to be made at a very low taxonomic level. In

Table 4. rRNA operon copy number estimated from metagenomic data gathered by ecosystem type (lake, estuary, coastal, open ocean), or unified (aquatic), compared with data obtained from the ribosomal RNA database (rrnDB)

	Lake	Estuary		Coastal		Open ocean		
	GS20	GS11	GS12	GS8	GS19	GS22	Aquatic	rrnaDB
Actinobacteria	1.5	1.5	1.3	6.5	2.3	4.2	2.9	3.1
Bacteroides	2.7	2.8	2.0	8.4	1.5	1.9	3.2	3.7
Cyanobacteria					1.8	1.8	1.8	2.4
Firmicutes	0.8						0.8	6.4
Gemmatimonas	2.0						2.0	nd
Green nonsulfur bacteria	1.9	3.1	1.3				2.1	1.7
Planctomyces	1.3						1.3	2
Proteobacteria	1.2	1.7	1.4	3.7	2.0	1.8	2.0	4.1
Alpha	1.5	2.0	1.1	5.9	1.9	1.8	2.3	2.4
Beta	1.0	2.3	4.3	3.0			2.7	3.9
Gamma	1.1	0.8	0.8	2.6	3.8	2.1	1.9	5.8
Verrucomicrobia	1.6	0.5		0.5			0.9	1.7
Archaea	1.8						1.8	1.8
Crenarchaeota	2.0						2.0	1

contrast, a phylogenetic tree affiliation is able to distinguish between sequences clearly related to a specific group with few references, and sequences presenting similarities with a set of unrelated taxa. Therefore, to affiliate reads more accurately, an assignment method based on phylogenetic trees allows the detection of distant homology for metagenomic fragments. In a previous work, Wu & Eisen (2008) developed an automatic pipeline for analyzing protein markers from assemblies, but their method might underestimate organisms with low depth coverage because of the reduction of the sample to a set of summarized information and assemblies which could lead to chimeric sequences. The main difference with MLTREEMAP (Stark *et al.*, 2010) is the set of markers used and the distinction between each protein marker, as MLTREEMAP results correspond to a concatenation of phylogenetic markers. The phylogenetic procedure implemented here is specific because of the limited size of reads to be affiliated and because these reads often only recover part of the reference sequences.

The use of the set of protein markers selected in this study offers added advantages, as these markers can assist in evolutionary inferences by increasing the number of informative characters that can be analyzed from the metagenome (Santos & Ochman, 2004). The affiliation of reads associated with *Methylophylales* (*Betaproteobacteria*) is a good example: some markers detect this group, whereas others consider the sequences 'unclassified'. Thus, the original bacterial group of the metagenomic sequences is likely to be related to *Methylophylales*, but is notably distinct from it. The same conclusions can be drawn from sequences related to the class *Actinobacteria*, which can be found in aquatic environments. Many protein-coding sequences emerged at the root of the *Actinobacteria* lineage when compared to organisms present in KEGG, reflecting their remote relatedness to them and the low representation of environmental species within sequenced genomes. Similar conclusions were reached by Philosof *et al.* (2009) analyzing three fosmids affiliated to *Actinobacteria* from Lake Kinneret and comparing the phylogeny of both 16S rRNA and pyrG genes physically linked on the same sequence. The pyrG phylogeny of one of these fosmids, affiliated to the acIV clade, indeed placed this sequence at the root of the actinobacterial lineage. This might indicate the major limitation of using protein marker genes, the limited number of reference sequences compared to 16S rRNA gene sequences. Nevertheless, as PCR primers (Santos & Ochman, 2004) are available for these protein-coding marker genes, the number of reference sequences could be easily and quickly increased.

The rRNA operon copy numbers, which are unknown for uncultured organisms (Sipos *et al.*, 2007), might also reflect differences between the species from databases and microorganisms in environments. This average estimate is essentially lower than the one in sequenced genomes, even for the dominant groups such as *Actinobacteria* and *Betaproteobacteria* in lakes or *Alphaproteobacteria* in marine environments and, more particularly, for members of *Firmicutes*, *Gammaproteobacteria* or *Crenarchaeota*. For instance, for this last phyla, the number of 16S rRNA gene copies is based on only one single genome among the mesophilic *Archaea* belonging to the Marine Archaeal Group I, *Nitrosopumilus maritimus* (Konneke *et al.*, 2005), whereas numerous environmental clades has been found (Schleper *et al.*, 2005).

Protein-driven analysis of the structure of bacterial and archaeal communities

Looking at the major groups found in aquatic ecosystems, e.g. *Actinobacteria*, *Bacteroides*, *Proteobacteria* and *Cyanobacteria*, there is a good agreement between the protein marker chosen and the results given by the 16S rRNA gene amplified or present in these metagenomes. Thus, *Actinobacteria* and *Betaproteobacteria* commonly are the main taxonomic groups found in lakes by PCR or metagenomic methods (Hahn, 2006), whereas coastal and ocean environments are dominated by *Alphaproteobacteria* (e.g. Pommier *et al.*, 2006). More precisely, typical freshwater or marine *Bacteria* such as *Burkholderiales* or *Pelagibacter* were recovered by protein markers when the information was available in the databases. *Pelagibacter* is thereby the major genus among clade SAR11 (Morris *et al.*, 2002), and the main typical freshwater clades (*Polynucleobacter* and R-BT065; Hahn, 2006) belong to the *Burkholderiales* order. Finally, *Alphaproteobacteria* communities in GS08 were not dominated by members of the genus *Pelagibacter* and the rRNA operon copy number estimation confirmed the specificity of bacterial communities from this sample. Compared to the other aquatic ecosystems, the estimated 16S rRNA gene copy number is slightly higher in GS08 for *Proteobacteria*, *Bacteroides* and *Actinobacteria*. According to previous work based on genome analysis, this copy number could be linked to carbohydrate transport and metabolism (Konstantinidis & Tiedje, 2004) and bacteria responding quickly to substrate availability (Klappenbach *et al.*, 2000). Like the specificity highlighted for *Alphaproteobacteria*, these differences in 16S rRNA operon copy number could be linked to the human impact on this ecosystem, as the GS08 sample came from Newport Harbor, a coastal station in the New England shelf region of the Mid-Atlantic Bight (Rusch *et al.*, 2007), and the only temperate coastal sample in our GOS subset. However, based on sample similarities calculated by Rusch *et al.* (2007), the GS08 sample is gathered with other coastal samples from the North

Atlantic coast, indicating that this could be a more general phenomenon.

A surprising result is the higher diversity highlighted by protein markers compared to 16S rRNA genes found in metagenomes or clone libraries. However, a false phylogenetic affiliation may be suspected when a taxonomic group is detected by only one phylogenetic marker in the absence of 16S rRNA genes in the metagenome reads, such as for *Acidobacteria* in GS08. On the other hand, the same group was PCR-amplified in the GS11 metagenome, despite only being detected by one protein marker. In addition, the taxonomy units detected by protein markers and not by 16S rRNA genes corresponded to groups rarely inventoried in aquatic ecosystems based on 16S rRNA genes. The most likely explanation is therefore that the sequencing effort was not detailed enough to capture these taxa. For example, *Acidobacteria* was mainly detected in soils or marine sediments but very rarely in the water column and only after a high-throughput identification of rRNA gene-containing clones in a large insert of a marine metagenomic library (Pham *et al.*, 2008). Similarly, although *Dictyoglomi* is a thermophilic bacteria, some sequences were also detected in clone libraries from mesophilic environments in a study dealing with OP11 clade (Harris *et al.*, 2004). Other groups represented by at least one protein marker were more abundant, such as *Archaea* or *Firmicutes*. The presence of *Archaea* in mesophilic marine environments has been established since 1992 (DeLong, 1992), whereas evidence of their abundance in lakes is more recent (Keough *et al.*, 2003). *Firmicutes* do not belong to the core species in lakes or oceans but have sometimes been detected in large clone libraries from lakes (Eiler & Bertilsson, 2004). Our analysis shows that *Firmicutes* could represent a significant component of lake bacterioplankton.

PCR amplification underestimates bacterial richness compared to protein-coding genes

Some cases revealed discrepancies between results obtained from metagenomic reads and from PCR amplifications from the same samples (Shaw *et al.*, 2008). PCR bias is a widely recognized phenomenon and affects major and minor groups in this study. For example, in GS11, it is clear that *Betaproteobacteria* were over-represented and *Actinobacteria* underestimated after amplification. In addition, some groups undetected by four protein markers were retrieved in PCR-amplified libraries. Finally, in terms of diversity, of 23 bacterial taxomomic units considered in this study, an average 10.7 were identified by PCR regardless of the ecosystem studied, compared with 21.7 detected by at least one phylogenetic marker in metagenomes (protein or 16S). These data can be compared to the study of Hong *et al.* (2009) who claimed that probably half of rRNA bacterial diversity is missed by PCR protocols.

The reported biases in PCR methods for studying microbial diversity are the specificity of primers targeting 16S rRNA gene, the number of 16S rRNA gene copies per cell, artefactual intraspecific sequence variation, inhibition of PCR by contaminants, and DNA extraction (Sipos *et al.*, 2007; Hong *et al.*, 2009; Morales & Holben, 2009). The DNA extraction used by Rusch *et al.* (2007) allowed the detection of a great variety of *Archaea* and *Bacteria* and does not seem a limiting factor in aquatic ecosystems, which is probably not the case in soils and sediments (Hong *et al.*, 2009). In addition, there was a significant correlation between the total rRNA gene copy number (rrna operons in Table 4 $\times$ abundances of taxon considered) and the abundance of amplified 16S rRNA ($r = 0.89$), according to common assumptions about the pitfalls of PCR (Farrelly *et al.*, 1995).

The specificity of primers associated with the sequencing effort is probably the critical step in retrieving accurate diversity data. For Armougom & Raoult (2009), the use of primers is undoubtedly one of the most critical factors affecting 16S rRNA gene analysis, and other studies highlighted that some primers are highly specific for a spectrum of bacterial species (Wang & Qian, 2009). Important members of a community may be overlooked when one mismatch is present with a primer and when stringent annealing temperatures are used (Sipos *et al.*, 2007). More to the point, universal primers failed to provide amplification products from 20% to 50% of *Actinobacteria* (Farris & Olson, 2007), which probably explains the differences observed with phylogenetic markers in metagenomic reads in GS11 and GS20. The primer set used for generating PCR libraries in these aquatic ecosystems (Shaw *et al.*, 2008), 27f-1492r, is commonly used in microbial ecology and has been critically evaluated (Frank *et al.*, 2008). These primers could preferentially amplify taxa such as *Alphaproteobacteria* and also detect microorganisms at low 16S rRNA gene copies such as TG1 (no 16S rRNA was detected in metagenomic reads) but miss more abundant taxa such as *Firmicutes*.

Whatever the reasons for this PCR bias, the global picture of diversity seen through protein markers and 16S rRNA genes in metagenomic reads, appears very dissimilar to the picture characterized by amplifying 16S rRNA genes. The structure of microbial assemblages and their variation in space and time cannot be assessed without a reliable method for evaluating the richness and diversity. Using a set of different markers could lead to the conclusion that 'everything is everywhere', according to the highly controversial debate (e.g. Martiny *et al.*, 2006), whereas from the PCR-mediated view, a biogeography of bacteria could be

deduced at the phylogenetic level presented. The use of different 16S primer sets or group-specific primers (Muhling *et al.*, 2008) could be a solution for in-depth biodiversity analysis. However, whatever the method used for obtaining 16S rRNA gene sequences, this marker still appears limited based on studies on *Pseudomonas* spp. (Cho & Tiedje, 2000) or hyperthermophilic *Archaea* (Whitaker *et al.*, 2003). Furthermore, Whitaker *et al.* (2003) also stated that several protein-coding genes were necessary to obtain a good enough resolution to indicate endemic clades. According to our study, exploring metagenomic data is a more reliable and less biased method than the PCR approach to study microbe diversity.

Finally, results presented here showed that a set of five efficient markers for which primers are available (Santos & Ochman, 2004) can be used for assessing microorganism diversity, by testing their phylogenetic resolution on real microbial assemblages, i.e. metagenomic data. However, metagenomic studies are known to be significantly more expensive than PCR approaches. Thus, PCR amplification of a set of protein-coding marker genes would make it possible to generate reliable phylogenies and, although not avoiding all the pitfalls of PCR, would limit the bias caused by the multiple copies or rRNA in a cell. Nevertheless, differences in codon usage occurring in the evolution of protein-coding genes could be a source of quantitative bias for PCR study of protein markers. These markers should be useful in the study of complex microbial communities as a complement to the information provided by 16S rRNA gene.

References

Amann RI, Ludwig W & Schleifer KH (1995) Phylogenetic identification and *in situ* detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143–169.

Armougom F & Raoult D (2009) Exploring microbial diversity using 16S rRNA high-throughput methods. *J Comput Sci Syst Biol* **02**: 74–92.

Bateman A, Birney E, Durbin R, Eddy SR, Finn RD & Sonnhammer EL (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* **27**: 260–262.

Biers EJ, Sun S & Howard EC (2009) Prokaryotic genomes and diversity in surface ocean waters: interrogating the global ocean sampling metagenome. *Appl Environ Microbiol* **75**: 2221–2229.

Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF & Kjelleberg S (2007) Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol* **73**: 278–288.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: 540–552.

Cho J-C & Tiedje JM (2000) Biogeography and degree of endemicity of fluorescent *Pseudomonas* strains in soil. *Appl Environ Microbiol* **66**: 5448–5456.

Coenye T & Vandamme P (2003) Intragenomic heterogeneity between multiple 16S ribosomal RNA operons in sequenced bacterial genomes. *FEMS Microbiol Lett* **228**: 45–49.

Dahllöf I, Baillie H & Kjelleberg S (2000) rpoB-based microbial community analysis avoids limitations inherent in 16S rRNA gene intraspecies heterogeneity. *Appl Environ Microbiol* **66**: 3376–3380.

Debroas D, Humbert JF, Enault F, Bronner G, Faubladier M & Cornillot E (2009) Metagenomic approach studying the taxonomic and functional diversity of the bacterial community in a mesotrophic lake (Lac du Bourget–France). *Environ Microbiol* **11**: 2412–2424.

DeLong EF (1992) Archaea in coastal marine environments. *P Natl Acad Sci USA* **89**: 5685–5689.

DeLong EF (1998) Everything in moderation: archaea as 'non-extremophiles'. *Curr Opin Genet Dev* **8**: 649–654.

Eddy SR (1998) Profile hidden Markov models. *Bioinformatics* **14**: 755–763.

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.

Eiler A & Bertilsson S (2004) Composition of freshwater bacterial communities associated with cyanobacterial blooms in four Swedish lakes. *Environ Microbiol* **6**: 1228–1243.

Farrelly V, Rainey F & Stackebrandt E (1995) Effect of genome size and rrn gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microbiol* **61**: 2798–2801.

Farris MH & Olson JB (2007) Detection of *Actinobacteria* cultivated from environmental samples reveals bias in universal primers. *Lett Appl Microbiol* **45**: 376–381.

Frank JA, Reich CI, Sharma S, Weisbaum JS, Wilson BA & Olsen GJ (2008) Critical Evaluation of two primers commonly used for amplification of bacterial 16S rRNA genes. *Appl Environ Microbiol* **74**: 2461–2470.

Giovannoni SJ & Stingl U (2005) Molecular diversity and ecology of microbial plankton. *Nature* **437**: 343–348.

Guindon S & Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: 696–704.

Hahn MW (2006) The microbial diversity of inland waters. *Curr Opin Biotechnol* **17**: 256–261.

Harris JK, Kelley ST & Pace NR (2004) New perspective on uncultured bacterial phylogenetic division OP11. *Appl Environ Microbiol* **70**: 845–849.

Hong S, Bunge J, Leslin C, Jeon S & Epstein SS (2009) Polymerase chain reaction primers miss half of rRNA microbial diversity. *ISME J* **3**: 1365–1373.

Huang Y, Gilna P & Li W (2009) Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics* **25**: 1338–1340.

Hugenholtz P (2002) Exploring prokaryotic diversity in the genomic era. *Genome Biol* **3**: REVIEWS0003.

Huson DH, Richter DC, Mitra S, Auch AF & Schuster SC (2009) Methods for comparative metagenomics. *BMC Bioinformatics* **10** (suppl 1): S12.

Jones DT, Taylor WR & Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**: 275–282.

Kanehisa M (2002) The KEGG database. *Novartis Found Symp* **247**: 91–101; discussion 101–103, 119–128, 244–252.

Keough BP, Schmidt TM & Hicks RE (2003) Archaeal nucleic acids in picoplankton from great lakes on three continents. *Microb Ecol* **46**: 238–248.

Klappenbach JA, Dunbar JM & Schmidt TM (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol* **66**: 1328–1333.

Konneke M, Bernhard AE, de la Torre JR, Walker CB, Waterbury JB & Stahl DA (2005) Isolation of an autotrophic ammonia-oxidizing marine archaeon. *Nature* **437**: 543–546.

Konstantinidis KT & Tiedje JM (2004) Trends between gene content and genome size in prokaryotic species with larger genomes. *P Natl Acad Sci U S A* **101**: 3160–3165.

Konstantinidis KT, Ramette A & Tiedje JM (2006) Toward a more robust assessment of intraspecies diversity, using fewer genetic markers. *Appl Environ Microbiol* **72**: 7286–7293.

Koski LB & Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* **52**: 540–542.

Ludwig W, Strunk O, Westram R *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.

Martiny JBH, Bohannan BJM, Brown JH *et al.* (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102–112.

Morales SE & Holben WE (2009) Empirical testing of 16S rRNA gene PCR primer pairs reveals variance in target specificity and efficacy not suggested by *in silico* analysis. *Appl Environ Microbiol* **75**: 2677–2683.

Morris RM, Rappe MS, Connon SA, Vergin KL, Siebold WA, Carlson CA & Giovannoni SJ (2002) SAR11 clade dominates ocean surface bacterioplankton communities. *Nature* **420**: 806–810.

Muhling M, Woolven-Allen J, Murrell JC & Joint I (2008) Improved group-specific PCR primers for denaturing gradient gel electrophoresis analysis of the genetic diversity of complex microbial communities. *ISME J* **2**: 379–392.

Pham VD, Konstantinidis KT, Palden T & DeLong EF (2008) Phylogenetic analyses of ribosomal DNA-containing bacterioplankton genome fragments from a 4000 m vertical profile in the North Pacific Subtropical Gyre. *Environ Microbiol* **10**: 2313–2330.

Philosof A, Sabehi G & Béjà O (2009) Comparative analyses of actinobacterial genomic fragments from Lake Kinneret. *Environ Microbiol* **11**: 3189–3200.

Pommier T, Canbäck B, Riemann L, Boström KH, Simu K, Lundberg P, Tunlid A & Hagström A (2006) Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* **16**: 867–880.

Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J & Glöckner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Rusch DB, Halpern AL, Sutton G *et al.* (2007) The sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. *PLoS Biol* **5**: e77.

Santos SR & Ochman H (2004) Identification and phylogenetic sorting of bacterial lineages with universally conserved genes and proteins. *Environ Microbiol* **6**: 754–759.

Schellenberg J, Links MG, Hill JE, Dumonceaux TJ, Peters GA, Tyler S, Ball TB, Severini A & Plummer FA (2009) Pyrosequencing of the chaperonin-60 universal target as a tool for determining microbial community composition. *Appl Environ Microbiol* **75**: 2889–2898.

Schleper C, Jurgens G & Jonuscheit M (2005) Genomic studies of uncultivated archaea. *Nat Rev Microbiol* **3**: 479–488.

Shaw AK, Halpern AL, Beeson K, Tran B, Venter JC & Martiny JBH (2008) It's all relative: ranking the diversity of aquatic bacterial communities. *Environ Microbiol* **10**: 2200–2210.

Sipos R, Székely AJ, Palatinszky M, Révész S, Márialigeti K & Nikolausz M (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *FEMS Microbiol Ecol* **60**: 341–350.

Stark M, Berger S, Stamatakis A & von Mering C (2010) MLTreeMap - accurate maximum likelihood placement of environmental DNA sequences into taxonomic and functional reference phylogenies. *BMC Genomics* **11**: 461.

Thornburg CC, Zabriskie TM & McPhail KL (2010) Deep-sea hydrothermal vents: potential hot spots for natural products discovery? *J Nat Prod* **73**: 489–499.

Venter JC, Remington K, Heidelberg JF *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**: 66–74.

von Mering C, Hugenholtz P, Raes J, Tringe SG, Doerks T, Jensen LJ, Ward N & Bork P (2007) Quantitative phylogenetic assessment of microbial communities in diverse environments. *Science* **315**: 1126–1130.

Wang Y & Qian P-Y (2009) Conservative fragments in bacterial 16S rRNA genes and primer design for 16S ribosomal DNA amplicons in metagenomic studies. *PLoS ONE* **4**: e7401.

Whitaker RJ, Grogan DW & Taylor JW (2003) Geographic barriers isolate endemic populations of hyperthermophilic Archaea. *Science* **301**: 976–978.

Wu M & Eisen JA (2008) A simple, fast, and accurate method of phylogenomic inference. *Genome Biol* **9**: R151.

Wu X, Monchy S, Taghavi S, Zhu W, Ramos J & van der Lelie D (2011) Comparative genomics and functional analysis of niche-specific adaptation in Pseudomonas putida. *FEMS Microbiol Rev* **35**: 299–323.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Schematic representation of the phylogenetic tree affiliation pipeline.
Fig. S2. Taxonomic affiliation for GS12 and GS08 ecosystems, with five protein markers (rplB, pyrG, leuS, fusA and rpoB), metagenomic 16S rRNA and PCR-amplified 16S rRNA.
Fig. S3. Taxonomic affiliation for minor groups (i.e. all groups with the exception of *Actinobacteria* and *Proteobacteria*) with five protein markers (rplB, pyrG, leuS, fusA and rpoB), metagenomic 16S rRNA and PCR-amplified 16S rRNA.
Table S1. Main characteristics of protein-coding genes used as phylogenetic markers.
Table S2. Percentage of correct affiliations for tree affiliation and best BLAST hit obtained from simulated data.
Table S3. *Alpha-Proteobacteria* affiliations for GS08.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

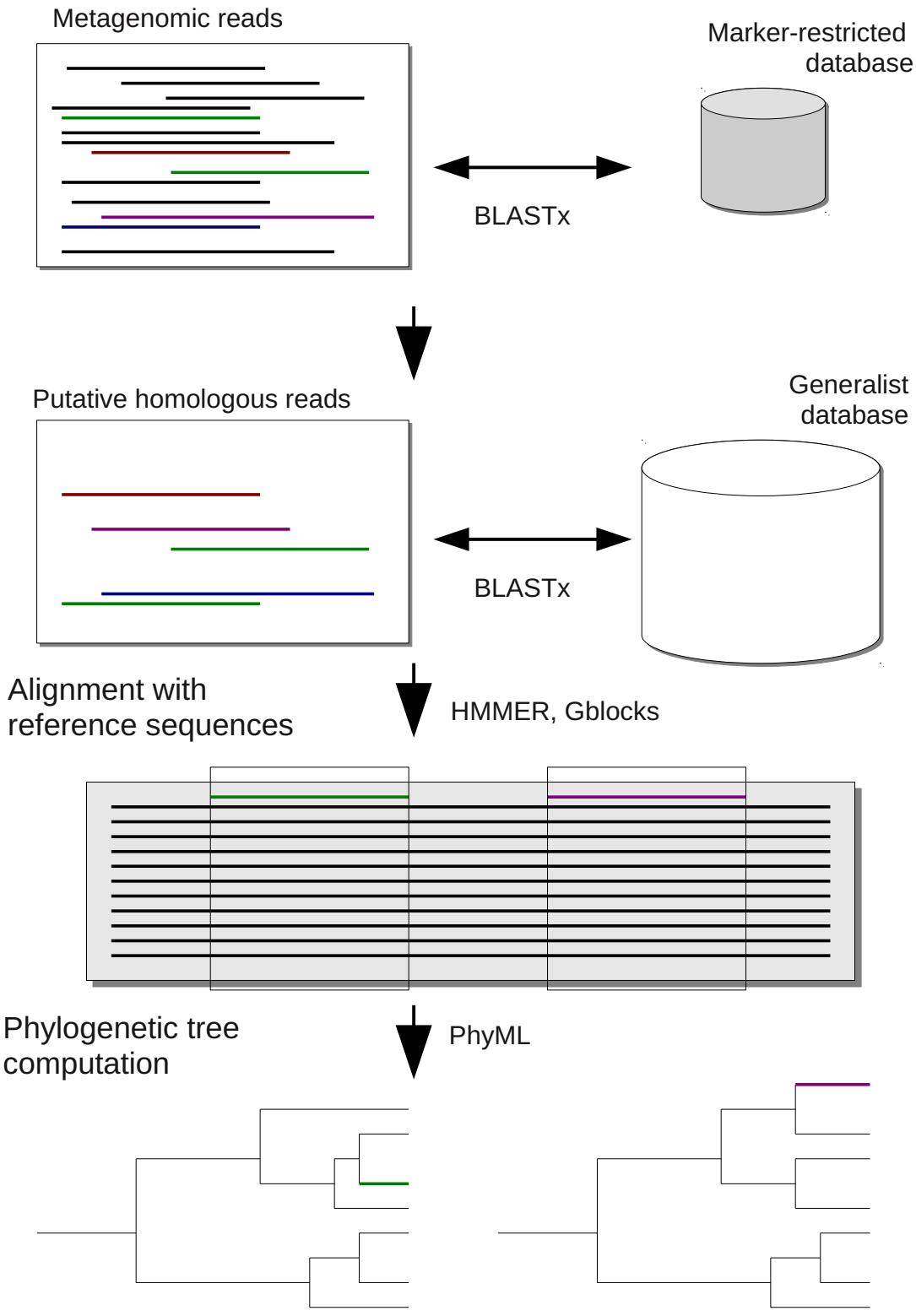

Metagenomic reads
Marker-restricted database
BLASTx
Putative homologous reads
Generalist database
BLASTx
Alignment with reference sequences
HMMER, Gblocks
Phylogenetic tree computation
PhyML

Fig S1

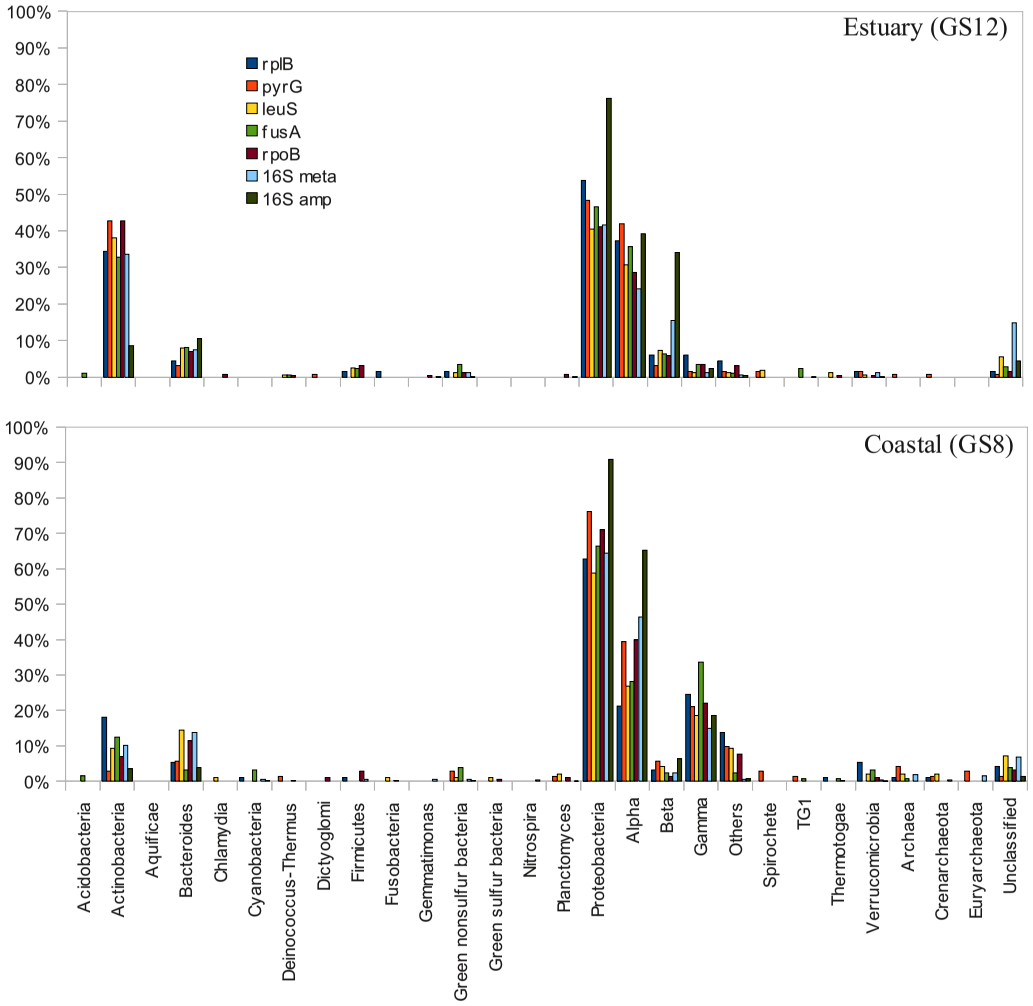

Estuary (GS12)
rplB
pyrG
leuS
fusA
rpoB
16S meta
16S amp
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
Coastal (GS8)
Acidobacteria
Actinobacteria
Aquificae
Bacteroides
Chlamydia
Cyanobacteria
Deinococcus-Thermus
Dictyoglomi
Firmicutes
Fusobacteria
Gemmatimonas
Green nonsulfur bacteria
Green sulfur bacteria
Nitrospira
Planctomyces
Proteobacteria
Alpha
Beta
Gamma
Others
Spirochete
TG1
Thermotogae
Verrucomicrobia
Archaea
Crenarchaeota
Euryarchaeota
Unclassified

Fig. S2

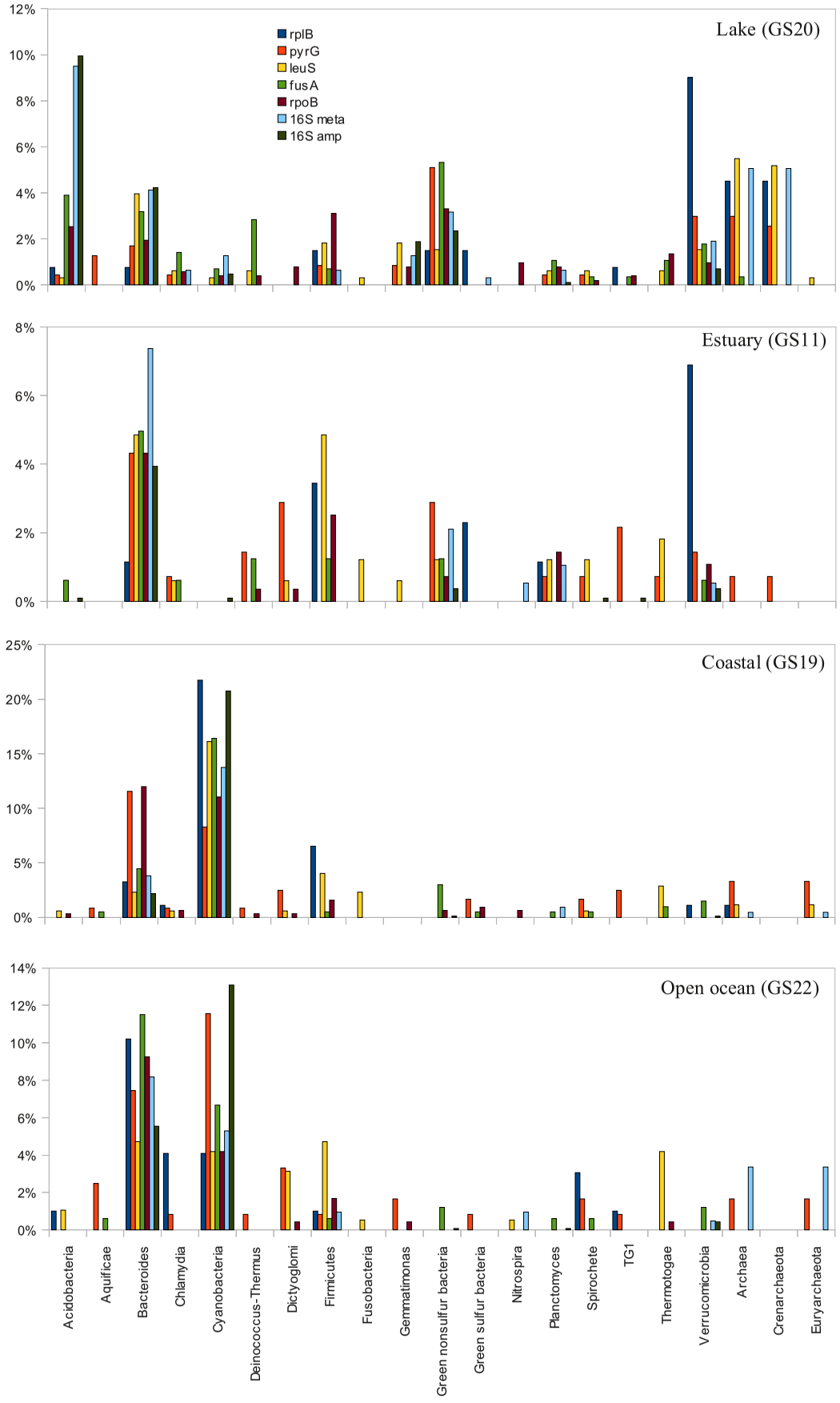

Lake (GS20)
rplB
pyrG
leuS
fusA
rpoB
16S meta
16S amp
12%
10%
8%
6%
4%
2%
0%
Estuary (GS11)
8%
6%
4%
2%
0%
Coastal (GS19)
25%
20%
15%
10%
5%
0%
Open ocean (GS22)
14%
12%
10%
8%
6%
4%
2%
0%
Acidobacteria
Aquificae
Bacteroides
Chlamydia
Cyanobacteria
Deinococcus-Thermus
Dictyoglomi
Firmicutes
Fusobacteria
Gemmatimonas
Green nonsulfur bacteria
Green sulfur bacteria
Nitrospira
Planctomyces
Spirochete
TG1
Thermotogae
Verrucomicrobia
Archaea
Crenarchaeota
Euryarchaeota

Fig. S3

Marker	Complete name	Sequence size (aa)	Alignment size
rplB	ribosomal protein L2	275	370
pyrG	CTP synthetase	550	840
fusA	Ef-G elongation factor	700	1545
leuS	leucyl t-RNA synthetase	860	2270
rpoB	RNA polymerase, β sub-unit	1500	4350

Table S1: Main characteristics of protein-coding genes used as phylogenetic markers.

Marker	Tree affiliation	Best blast hit affiliation
rplB	99.20%	95.00%
pyrG	98.60%	91.40%
fusA	93.76%	94.00%
leuS	94.76%	90.60%
rpoB	98.10%	86.20%

Table S2: Percentage of correct affiliations for tree affiliation and best BLAST hit obtained from simulated data. The tree affiliation ratio is calculated as the ratio between correct affiliation and the total number of tree generated. Best BLAST hit rate is the number of correct affiliation on the total number of sequences studied.

	rplB	pyrG	leuS	fusA	rpoB	16S meta	16S amp
Jannaschia	-	-	-	16.7%	2.7%	-	-
Magnetospirillum	10.0%	7.1%	7.7%	-	0.7%	-	-
Others	25.0%	57.1%	42.3%	66.7%	32.7%	15.5%	7.3%
Pelagibacter	40.0%	14.3%	30.8%	8.3%	60.0%	67.7%	17.3%
Rhodobacter	-	-	-	-	-	11.0%	58.6%
Rhodospirillum	15.0%	7.1%	11.5%	8.3%	0.7%	5.8%	16.8%
Roseobacter	10.0%	14.3%	-	-	3.3%	-	-
Sphingomonas	-	-	7.7%	-	-	-	-

Table S3: *Alpha-Proteobacteria* affiliations for GS08. "-" indicates that no sequences were affiliated to this group for this marker. "Others" correspond to sequences associated with *Alpha-Proteobacteria*, but not inserted in one of the groups studied here.

Annexe A.2 : Article

Application des approches métagénomiques à l'étude de la diversité virale environnementale.

Simon Roux[1,2], Didier Debroas[1,2], François Enault[1,2]

[1]Laboratoire Microorganismes: Génome et Environnement, Clermont Université, Université Blaise Pascal, BP 10448, F-63000 Clermont-Ferrand

[2]CNRS, UMR 6023, LMGE, F-63177 Aubière

[3]Centre Régional de Ressources Informatiques, Clermont Université, Université Blaise Pascal, Clermont-Ferrand, France

Paru dans **Virologie** (2013; 17(4) : 229-42)

revue

Virologie

Virologie 2013, 17 (4) : 229-42

Application des approches métagénomiques à l'étude de la diversité virale environnementale

Simon Roux[1,2]
Didier Debroas[1,2]
François Enault[1,2]

[1] Clermont université,
université Blaise-Pascal,
laboratoire micro-organismes : génome
et environnement,
24, avenue des Landais, 63171 Aubière,
France
[2] CNRS, UMR 6023,
laboratoire micro-organismes : génome
et environnement, 63171 Aubière,
France
<simon.roux@univ-bpclermont.fr>

Résumé. Les virus de l'environnement sont à la fois très nombreux, très diversifiés et largement méconnus. Au-delà de leur pouvoir pathogène, on leur reconnaît aujourd'hui une influence plus large sur des aspects fondamentaux de l'écologie de notre planète, comme les cycles biogéochimiques, la régulation des communautés de micro-organismes ou encore l'évolution des organismes vivants et de leurs génomes. Les approches de métagénomique virale, consistant en un séquençage aléatoire massif des acides nucléiques encapsidés, ont permis durant cette dernière décennie de mieux connaître la composition des communautés virales naturelles ainsi que la diversité génétique des virus. Les communautés virales étudiées jusqu'à présent sont très diversifiées et riches, le type d'environnement semblant constituer le principal facteur influençant la composition de ces communautés. De plus, les séquences des métagénomes viraux sont le plus souvent éloignées des séquences de référence disponibles, témoignant du fait que la grande majorité des virus de l'environnement nous est encore inconnue.

Mots clés : métagénomique, diversité, virus

Abstract. Viruses are the most abundant biological entities observed in environmental samples. They display an extraordinary morphological and genetic richness. In addition to their pathogenicity, viruses are now considered as of major influence on biogeochemical cycles, microorganisms' population regulation and more generally on the evolution of cellular genomes throughout the history of life. Viral metagenomics, *i.e.* the random sequencing of encapsidated nucleic acids in samples, has provided important new insights into viral diversity. These data reveals an overwhelming genetic diversity in the biosphere's viral communities sampled so far, especially in marine environments. Generally, the type of biome or environmental niche from which samples were obtained seems to be the main determinant of its composition. In addition, virome sequences obtained are generally quite distant from the reference sequences available in databases. This fact leads to the inescapable conclusion that nearly all of the virosphere's vast population is currently unknown to science, and emphasizes the need for pushing efforts toward surveying and characterizing the viral diversity.

Key words: metagenomics, diversity, virus

Introduction

Les virus sont considérés comme les entités biologiques les plus abondantes de la biosphère. La présence de virions a été relevée dans l'ensemble des écosystèmes étudiés jusqu'à maintenant, depuis les environnements associés à l'homme (microbiome humain, environnements contaminés ou impactés par l'activité humaine) aux milieux naturels (sol, milieux lacustres, océans) y compris les environnements les plus extrêmes (proche de la saturation en sel, fortement acide, à température élevée, etc.). À titre d'exemple, des concentrations de 10^8 particules virales par millilitre ont été mesurées dans différents prélèvements océaniques et on estime que les océans abriteraient 4.10^{30} virus, soit l'équivalent en carbone d'environ 75 millions de baleines bleues [1]. Les mesures de ces fortes abondances pourraient être biaisées par certaines

doi:10.1684/vir.2013.0506

Pour citer cet article : Roux S, Debroas D, Enault F. Application des approches métagénomiques à l'étude de la diversité virale environnementale. *Virologie* 2013; 17(4) : 229-42 doi:10.1684/vir.2013.0506

limitations méthodologiques [2], mais il n'en reste pas moins qu'il existe une diversité morphologique exceptionnelle au sein du monde viral, comme en témoignent les observations en microscopie électronique à transmission (MET). En l'état actuel des connaissances, les capsides de types icosahédriques restent la forme la plus couramment retrouvée, mais des observations récentes font état de virus en forme d'ampoule, de goutte ou encore de bacille *(figure 1)* [3, 4]. Afin de mieux caractériser ces virus, les approches d'isolement et de culture de souches virales sont utilisées et permettent d'avoir accès au génome complet des souches cultivées [5, 6]. Ces génomes viraux entièrement séquencés sont très variables en termes de support, taille et contenu. En effet, pas moins de sept supports génétiques différents ont été décrits pour les génomes viraux (ADN double et simple brins, ARN double brin, ARN simple brin sens et antisens, ARN rétrotranscrit et ADN rétrotranscrit) et ces génomes, circulaires ou linéaires, ont une taille variant de 1 680 à 1 259 000 paires de bases (pb). De plus, ils abritent un nombre très important de nouveaux gènes, pour lesquels aucune séquence similaire n'est encore décrite [7]. Cependant, la mise en culture d'un virus nécessite de cultiver l'organisme cellulaire hôte de ce virus, ce qui est encore impossible aujourd'hui pour la majorité des organismes, en particulier pour les micro-organismes. De fait, le nombre de virus cultivables reste très faible par rapport au nombre et à la diversité des virus observés.

Afin d'étudier les communautés virales environnementales, de nouvelles méthodologies contournant cette étape de mise en culture ont été développées *(figure 2)*. Historiquement, les premières analyses des communautés virales (au sens écologique du terme) ont été effectuées par l'intermédiaire d'observations au MET associées à des comptages [8, 9]. Si ces approches ont notamment permis de mieux décrire les différentes morphologies existantes, l'analyse reste limitée par l'existence de virus similaires du point de vue morphologique mais différents tant par leurs hôtes que dans le contenu de leur génome. Ainsi, différentes approches d'écologie moléculaires ont été utilisées par la suite afin d'accéder à la diversité génétique et génomique de ces virus. La communauté virale d'un écosystème peut, par exemple, être caractérisée par un profil génétique *via* l'utilisation d'électrophorèses sur gel en gradient dénaturant (DGGE) ou en champ pulsé (PFGE) [10]. Ces profils permettent de comparer plusieurs échantillons, mais ne donnent malheureusement pas accès à l'information génétique en tant que telle. L'étude de gènes d'intérêt par PCR puis séquençage peut fournir de précieuses informations sur la diversité de ces gènes et, par extension, des organismes qui abritent ces gènes. Appliquée à un gène comme celui codant pour l'ARN 16S, conservé chez tous les procaryotes, cette méthodologie permet d'appréhender la diversité de ces micro-organismes dans un écosystème donné. Des approches similaires ont été appliquées avec succès à certaines familles virales [11] mais ne peuvent être transposées aux virus dans leur ensemble, aucun gène n'étant conservé dans l'ensemble des génomes viraux. De plus, ce type d'analyse ne peut pas être utilisé dans le cadre de recherches exploratoires, puisqu'il est nécessaire de disposer préalablement de séquences de référence du gène marqueur choisi afin de pouvoir mettre au point les amorces PCR.

Dans ce contexte d'un monde viral à la fois très diversifié et peu caractérisé, la métagénomique s'impose comme une approche de choix, puisqu'elle s'affranchit des limites de la mise en culture ainsi que de la nécessité de connaissances préalables des autres méthodes traditionnelles. Cette approche vise à séquencer des fragments aléatoires de génomes viraux issus d'un échantillon d'intérêt. L'application de cette technique à des échantillons de différents écosystèmes a permis une meilleure compréhension de la composition des communautés virales de l'environnement.

Les premières études de métagénomique virale furent publiées au début des années 2000 [12, 13] et confirmèrent l'hypothèse d'une diversité virale majoritairement inconnue *(figure 2)*. Toutefois, ce n'est qu'avec l'avènement des nouvelles techniques de séquençage (*next generation sequencing* [NGS]), offrant un accès à une très grande quantité de données, qu'il a été possible de véritablement évaluer l'étendue de la diversité génétique des communautés virales de l'environnement [14]. Cette première étude de métagénomique virale en milieu océanique à l'aide des NGS mit en évidence le nombre très important d'espèces virales présentes dans l'échantillon (estimé à plusieurs centaines de milliers), ainsi que la présence importante de gènes issus de transfert avec l'hôte (dans ce cas des bactéries). Différents types d'environnements ont ensuite été étudiés suivant la même approche, depuis les milieux hypersalins [15] aux fèces animaux [16]. Ces études ont mis en avant la place importante des bactériophages (majoritairement du groupe des Caudovirales, phages bactériens à queue) au sein des communautés virales de l'environnement, ainsi que la présence importante de petits virus à ADN simple brin [17]. Ces derniers étaient peu pris en compte jusqu'ici car la taille de leur capside (entre 15 et 30 nm) rend difficile toute observation ou comptage en microscopie. La flore virale associée au microbiome intestinal humain a aussi pu être étudiée de manière plus exhaustive : une série d'échantillons prélevées sur des jumeaux monozygotes et leur mères a notamment révélé que si les communautés bactériennes d'individus génétiquement liés sont généralement similaires, les communautés virales étaient en revanche uniques pour chaque individu et relativement

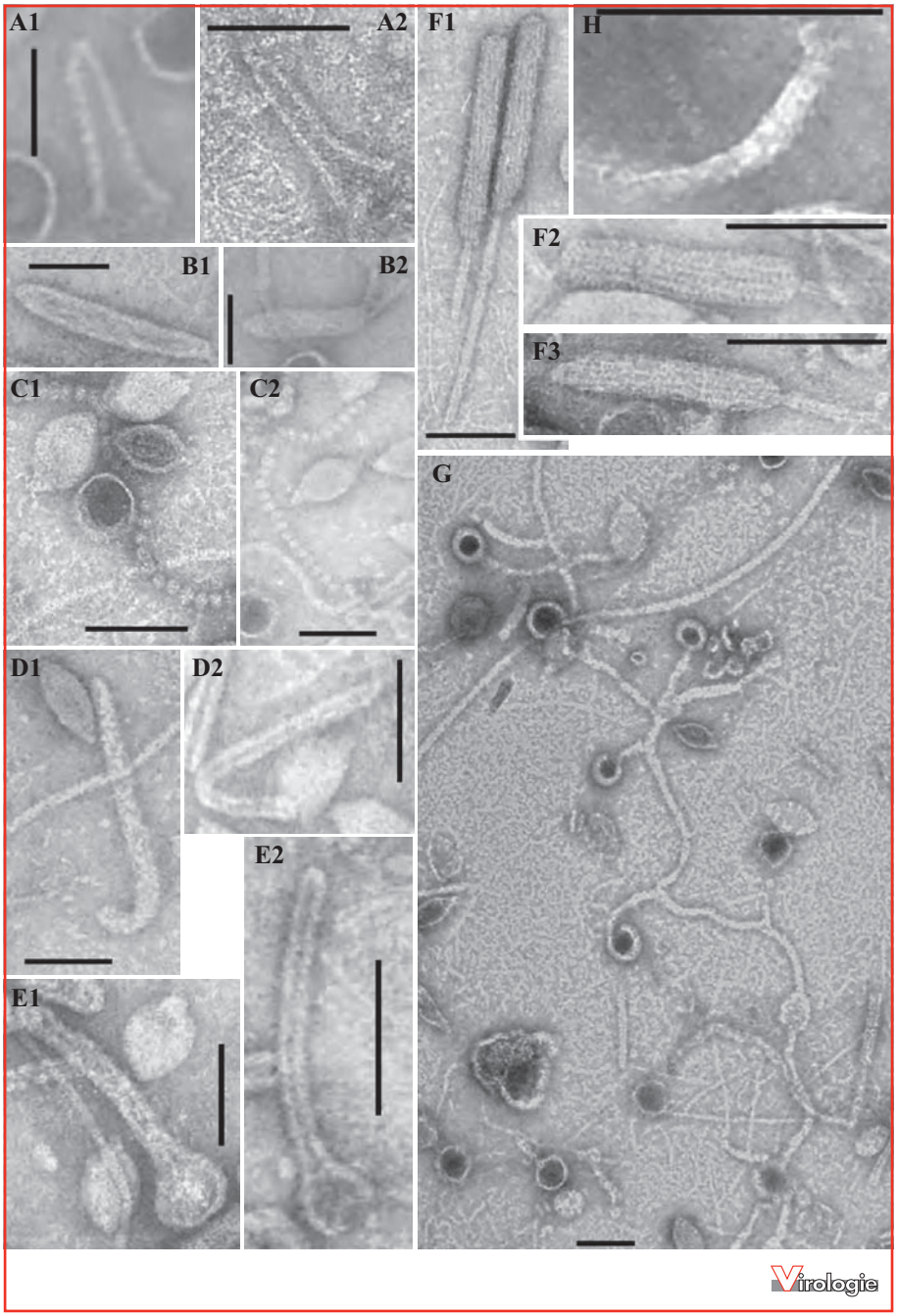


Figure 1. Particules virales de morphologies exceptionnelles observées au microscope électronique au sein d'échantillons du lac Rose (Sénégal). **A)** particule en forme d'épingle à cheveux ; **B)** particules en forme de bacille ; **C)** chaînes de globules ; **D)** particules en forme de crochet ; **E)** particules en forme de « têtard », formées par deux unités, l'une sphérique et l'autre linéaire ; **F)** particules en forme de « roseau », où des structures sphériques sont attachées au bout des branches ; **G)** particules complexes apparaissant comme un réseau de filaments connectés, associés à des structures sphériques aux extrémités ; **H)** structures terminales en forme de crochet, observées sur certaines particules. L'échelle représente 100 nm.
Figure issue de Sime-Ngando *et al.* (2010) [3].

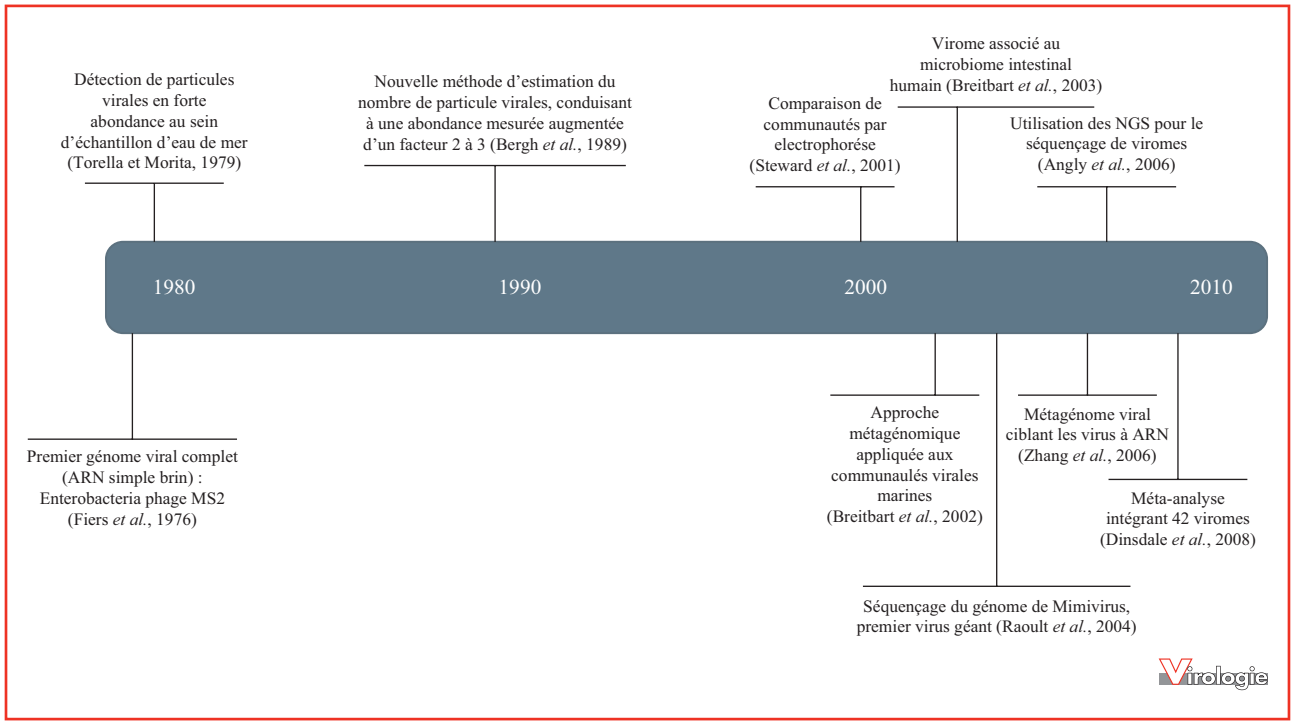


Figure 2. Vue d'ensemble des techniques d'études des communautés virales de l'environnement. Les premiers résultats étaient composés d'observations et de comptages. Les approches basées sur les techniques d'électrophorèse ont ensuite permis de comparer les communautés sur la base de profils basés sur une digestion des génomes. Le premier métagénome viral a été décrit en 2002 et mettait en lumière le caractère majoritairement inconnu des génomes viraux. En quelques années, le nombre d'études métagénomiques virales s'est multiplié, jusqu'à permettre la réalisation de méta-analyses en 2008.
Figure issue de Sime-Ngando *et al.* (2010) [3].

stables dans le temps [18]. Dans le but de mieux comprendre les relations et différences entre les virus d'environnements différents sur la planète, plusieurs méta-analyses ont été menées sur l'ensemble des données métagénomiques virales disponibles. Ces analyses ont mis en lumière le potentiel fonctionnel insoupçonné des communautés virales de l'environnement [19] et la spécificité importante de ces communautés par rapport aux types d'environnements échantillonnés [20]. Enfin, la grande majorité de ces études cible les séquences d'ADN encapsidés, et laisse de côté l'ensemble des virus à ARN. Ces derniers constitueraient pourtant une part importante des communautés virales de l'environnement [21] et, à ce titre, ne peuvent être ignorés lorsqu'il s'agit d'évaluer la diversité virale dans son ensemble.

Nous allons ici décrire les différentes étapes d'obtention et d'analyse de métagénomes viraux. Tout au long de cet article, l'analyse de deux viromes lacustres (issus de prélèvements des lacs Pavin et Bourget) sera utilisée comme un exemple des différentes analyses envisageables pour un virome [22]. Les résultats décrits, obtenus *via* le serveur d'analyse Metavir [23] développé au sein du laboratoire micro-organismes : génome et environnement (CNRS, UMR 6023), seront généralisés et comparés aux autres viromes publiés.

Isolement et séquençage de l'ADN viral

La construction d'un métagénome viral (ou virome) commence par une phase de préparation de l'échantillon, qui consiste à ne conserver que les virions et à en extraire les acides nucléiques. Cette étape est réalisée par une combinaison d'étapes de filtration afin d'éliminer la plus grande partie des cellules et de conserver la plupart des virus. Généralement, la fraction utilisée pour la préparation d'un virome est la fraction inférieure à 0,45 µm, voire inférieure à 0,2 µm. Il est toutefois à noter que de tels filtres excluent *de facto* les virus géants ou girus dont le représentant le plus célèbre est *Mimivirus*. La filtration en flux tangentiel (TFF) est ensuite souvent utilisée pour concentrer les particules virales en les retenant sur un filtre (généralement entre 30 et 100 kD) [24, 25]. Différentes alternatives existent pour

réaliser cette concentration, notamment basées sur une précipitation chimique des capsides virales (floculation au fer [26], précipitation au polyéthylène glycol [27]) ou sur une centrifugation [28]. Les capsides virales sont ensuite purifiées, afin de les séparer des matériaux et autres éléments également concentrés. Certains agents chimiques servent à la fois à la concentration et à la purification, comme le polyéthylène glycol. D'autres traitements chimiques ne permettent que de purifier les capsides virales, notamment les gradients de densité (de type chlorure de césium ou sucrose). Enfin, le matériel génétique libre doit être éliminé *via* différents traitements enzymatiques (*DNAse* et *RNAse*). L'ensemble de ces étapes permet d'obtenir un échantillon fortement enrichi en virus, voire même une purification totale de ces particules [24, 25]. Différentes techniques de séquençage peuvent alors être appliquées à ce matériel génétique isolé. Les premiers viromes publiés utilisaient la technique de Sanger, qui permet d'obtenir quelques milliers de fragments [12, 13]. L'apparition et l'évolution très rapide de NGS ont profondément modifié les données obtenues et les analyses qui en découlent *(figure 3)*. D'une manière générale, ces nouvelles techniques ont permis d'obtenir une profondeur de séquençage sans précédent, proposant dans un premier temps des centaines de milliers de séquences, jusqu'à plusieurs dizaine de millions aujourd'hui pour certaines méthodes comme l'Illumina Hi-Seq.

Les gènes des virus de l'environnement sont majoritairement inconnus

La première étape lors de l'analyse d'un virome est généralement d'essayer d'affilier chaque séquence à une séquence déjà connue et présente dans une base de données, de manière à déterminer la composition taxonomique et fonctionnelle de la communauté séquencée. Pour ce faire, les séquences métagénomiques sont comparées aux principales bases de données de séquences grâce à l'outil BLAST [29]. À de rares exceptions près, comme par exemple un virome

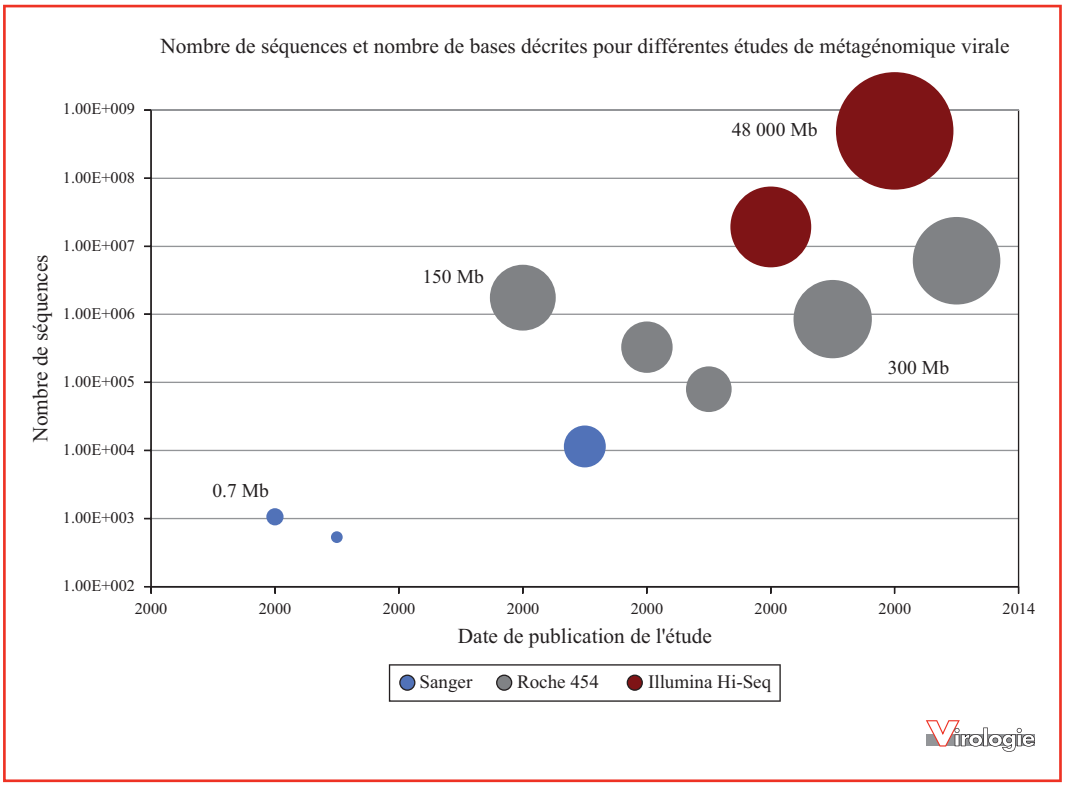


Figure 3. Graphique rapportant le nombre de séquences (en ordonnées) et le nombre de paires de bases (représenté par l'aire du cercle) pour différentes études de métagénomique virale. Le type de séquenceur utilisé est indiqué par un code couleur. Les valeurs de nombre de paires de bases sont indiquées pour quatre études représentatives des séquençage de type Sanger (Breitbart *et al.*, 2002 [12]), Roche 454 de première génération (GS20 ; Angly *et al.*, 2006 [14]), Roche 454 de troisième génération (titanium ; Minot *et al.*, 2011 [32]) et enfin Illumina Hi-Seq (Minot *et al.*, 2012 [44]).
Figure issue de Sime-Ngando *et al.* (2010) [3].

marin ciblant les virus à ARN [30] ou un virome issu d'un prélèvement atmosphérique [31], les métagénomes viraux présentent une très forte proportion de séquences inconnues (de 70 à 99 %, voir par exemple [14, 19, 22, 32]). La présence de séquences inconnues ne semble être ni liée à l'écosystème étudié ni aux méthodes utilisées, mais bien refléter une réalité générale quant à la part importante de gènes non caractérisés portés par les génomes viraux. Même si la faible taille des séquences de viromes influe sur ce taux de gènes non caractérisés, ce résultat est proche des taux de gènes non caractérisés observés lors du séquençage de nouveaux virus. Les deux viromes lacustres que nous avons caractérisés au laboratoire ne dérogent pas à la règle, seulement 10 et 20 % des séquences sont affiliées respectivement pour les lacs Pavin et Bourget. Ces taux sont plus faibles que ceux des viromes océaniques ou des viromes associés au microbiome humain et témoignent du manque de connaissance actuelle sur les communautés virales des environnements lacustres.

De plus, une proportion non négligeable des séquences affiliées sont similaires à des séquences issues de micro-organismes, principalement de bactéries. Plusieurs phénomènes peuvent expliquer la présence, contradictoire en apparence, de séquences issues de génomes d'organismes cellulaires dans des métagénomes viraux. En premier lieu, l'échange de gènes entre les génomes des virus et de leur hôte ainsi que l'intégration de génomes viraux au sein du génome de l'hôte sont des phénomènes bien documentés. Ainsi, certaines séquences de viromes sont bien des gènes viraux, pour lesquels les seuls exemplaires décrits sont des copies intégrées à des génomes cellulaires. Ces échanges, associés au faible nombre de génomes viraux de référence disponibles à l'heure actuelle, expliquent en partie ces similarités entre séquences de viromes et génomes cellulaires. Toutefois, certains cas semblent relever plus de la contamination du virome par le génome d'un organisme cellulaire, ces contaminations étant liées aux difficultés méthodologiques inhérentes à la conception d'un virome, en l'absence actuelle d'une standardisation des protocoles pour ce domaine en plein essor [24].

Les gènes viraux sont extrêmement diversifiés

S'il existe ainsi dans les viromes une quantité importante de séquences inconnues, il est important d'essayer de déterminer leur niveau de redondance et de diversité, à la fois au sein d'un échantillon et entre les différents types d'écosystème. Ainsi, l'étape suivante a pour objectif d'estimer la diversité génétique présente au sein de chaque jeu de données à partir de la redondance des séquences étudiées. Pour cela, les séquences similaires sont regroupées (*clusterisation*), le nombre de groupes créés et leur composition permettant ensuite d'estimer la richesse spécifique des génomes viraux, connues comme inconnues. La richesse en gènes des communautés virales des milieux aquatiques tempérés semble globalement élevée, particulièrement pour les milieux océaniques *(figure 4A)*. Les virus associés aux microbiomes eucaryotes (prélèvement de fèces, de salive ou de mucus pulmonaire sur des sujets humains, ou analyse d'échantillons issus de la dissection de poissons ou d'insectes) semblent eux présenter une richesse en gènes moins importante. De plus, les courbes de raréfaction, représentant de manière visuelle ces regroupements, n'atteignent que très rarement un plateau pour l'ensemble des viromes étudiés. Les séquences présentes dans les viromes sont donc loin de représenter la totalité des gènes viraux compris dans l'écosystème d'intérêt *(figure 5)*.

En parallèle de cette richesse génétique, l'outil PHACCS [33] permet d'estimer une richesse spécifique taxonomique, soit un nombre total de virotypes (ou espèces virales) différents contenus dans l'échantillon *(figure 4B)*. À l'inverse de la richesse en gènes, aucune différence entre les différents types d'échantillons n'apparaît. Ce résultat pourrait être lié à un biais méthodologique. En effet, l'outil PHACCS effectue une comparaison entre les résultats d'assemblage des séquences et des modélisations de ces mêmes assemblages selon différentes lois de distribution des espèces pour estimer les paramètres (type de distribution, nombre de génotypes) les plus plausibles. Or ces modélisations sont basées sur une taille moyenne de génomes estimée à partir des séquences affiliées, c'est-à-dire à partir d'une minorité des séquences de viromes. Toutefois, ces résultats contrastés pourraient également refléter les différences importantes de structures entre les populations virales de ces différents types d'échantillons, une dissociation pouvant exister entre la richesse génétique des communautés et le nombre de souches virales qui les composent.

Le type de milieu conditionne la composition des communautés virales

Les métagénomes viraux permettent de comparer les communautés virales et ainsi de mieux comprendre la répartition des populations virales dans la biosphère (ou bêta-diversité). La question des facteurs expliquant la répartition et la dispersion des populations virales est centrale pour la compréhension des rôles et impacts des communautés virales dans l'environnement, et reste encore ouverte [34, 35]. Dans le cas des métagénomes bactériens ou eucaryotes, une comparaison des affiliations des séquences est généralement réalisée. Toutefois, si ce type de comparaison

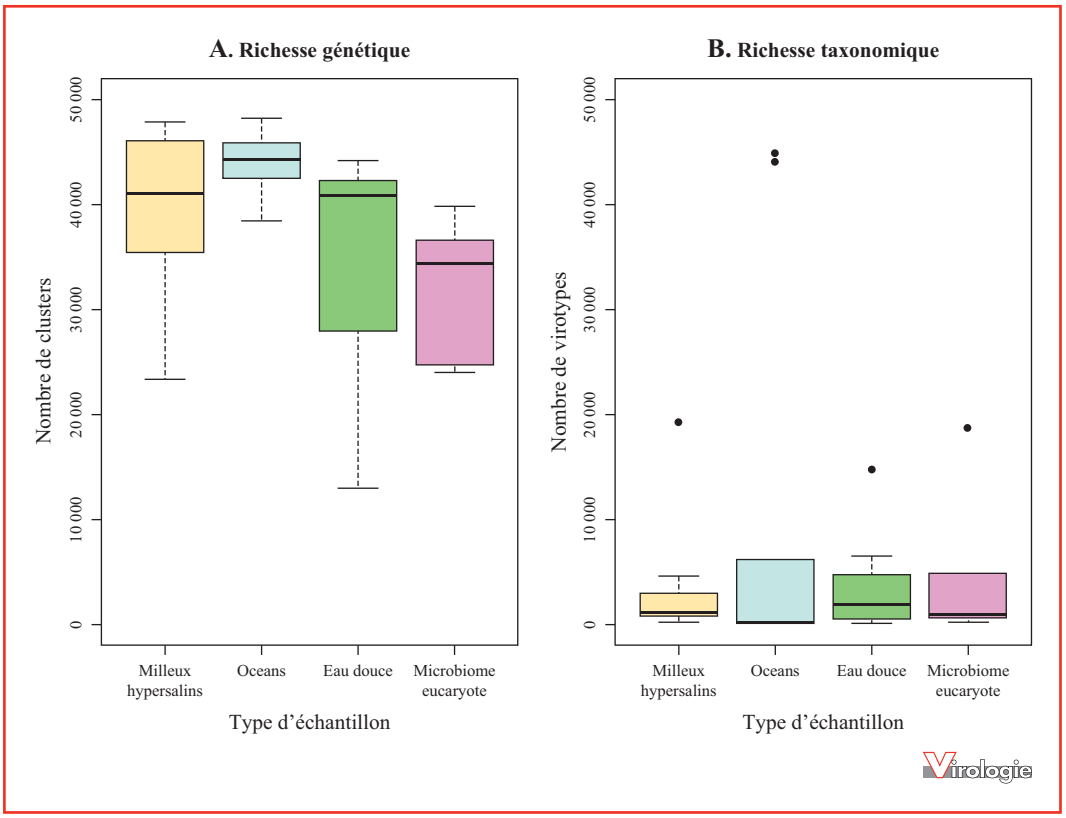


Figure 4. Distribution des richesses en gène et en espèce de viromes classés par environnement. **A)** la richesse en gène est estimée par le nombre de séquences différentes pour chaque virome pour un sous-échantillon de 50 000 séquences. Cette valeur reflète le nombre de gènes différents au sein des virus de l'échantillon ; **B)** la richesse en espèces correspond au nombre de virotypes présents dans l'échantillon, estimé par l'outil PHACCS.
Figure issue de Roux *et al.*, 2012 [22].

est applicable aux métagénomes pour lesquels la majorité des séquences est connue, il ne peut en aucun cas s'appliquer aux viromes où seule une petite fraction des jeux de données serait prise en compte. Afin de dépasser cette limite, il est possible de réaliser des comparaisons directes de l'ensemble des séquences des viromes. Ce type de comparaison ne permet pas d'expliquer quels sont les virus ou groupes de virus similaires entre plusieurs échantillons mais permet à tout le moins de révéler des similarités entre les communautés virales étudiées, qu'elles soient formées de virus connus ou inconnus. Les résultats de ces comparaisons montrent que les communautés virales issues du même type de prélèvement semblent présenter de fortes similarités en termes de contenu en gènes, et ce quelle que soit la distance géographique pouvant séparer les lieux de prélèvement *(figure 6)*. Cette comparaison permet tout d'abord de distinguer de manière nette les échantillons associés aux microbiomes eucaryotes des échantillons aquatiques. Au sein de ces deux catégories, les échantillons de même type sont à nouveau regroupés, avec, par exemple pour les échantillons aquatiques, une séparation entre environnements fortement hypersalins, environnements marins et faiblement hypersalins, et enfin les environnements d'eau douce. Ce type de distribution laisse entendre que les communautés virales seraient spécifiques de l'environnement dans lequel elles évoluent (potentiellement *via* les communautés d'hôtes spécifiques de ces environnements), et qu'il existerait des échanges entre ces communautés au niveau mondial.

Les viromes apportent une meilleure compréhension des principaux groupes viraux : exemple des Caudovirales, les bactériophages les plus abondants de la virosphère

Parmi les séquences de viromes affiliées aux virus, la majorité est similaire à des génomes de bactériophages,

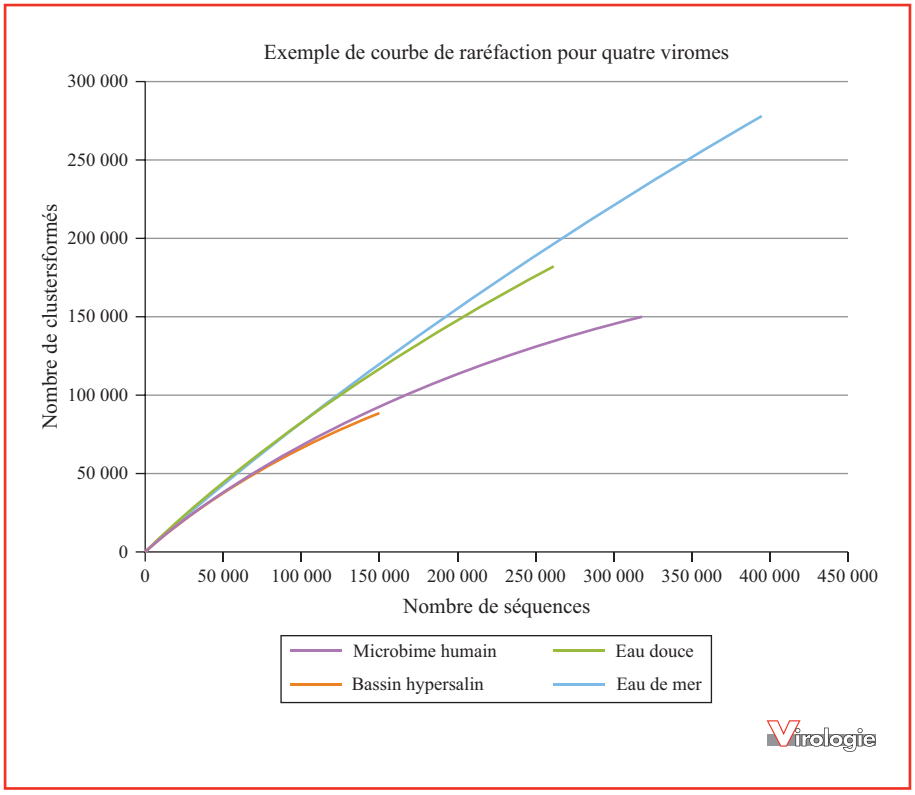


Figure 5. Courbes de raréfaction de viromes issus de quatre différents types d'échantillon. Pour chaque virome, la courbe représente le nombre de clusters (ou groupes de séquences similaires) différents obtenus en fonction du nombre de séquences du virome considéré. Figure issue de Roux *et al.*, 2012 [22].

appartenant principalement au groupe des Caudovirales (groupe de phages bactériens à structure tête-queue). Afin d'avoir une idée plus précise de la diversité de ces virus à partir des viromes, plusieurs types d'analyses sont possibles :

– la génération d'arbres phylogénétiques qui permettent de préciser la diversité ainsi que les liens évolutifs entre ces virus ;

– les graphiques de recrutement, qui permettent d'observer quels gènes du génome de référence sont retrouvés dans le métagénome.

L'existence de gènes conservés au sein des différents groupes viraux rend en effet possible de mener des analyses phylogénétiques pour ces groupes. Le gène codant pour la grande sous-unité de la terminase (*TerL*) est le marqueur le plus utilisé dans le cas des Caudovirales. Les phylogénies réalisées à partir des séquences de *TerL* retrouvées dans les viromes des lacs Pavin et Bourget ont mis à jour une grande diversité de ces virus dans ces lacs *(figure 7)*. En effet, les séquences métagénomiques sont distribuées sur l'ensemble de l'arbre et sont, pour la plupart, loin de toute séquence de référence. Ce résultat illustre à la fois l'existence de nombreux clades existant au sein des Caudovirales, et les différences importantes entre ces clades environnementaux et les séquences des bases de données.

Une partie de ces séquences de viromes sont affiliées au groupe des phages de type T4, groupe de référence incluant le phage T4 infectant la bactérie *Escherichia coli* et utilisé comme modèle depuis plusieurs décennies. Deux sous-groupes sont ainsi retrouvés dans les viromes lacustres : le groupe des cyanophages de type T4 et le groupe des Far-T4 [36], pour lequel le phage « *Rhodothermus* phage RM378 » est le génome de référence le plus proche *(figure 7)*. Si la présence de cyanophages dans des échantillons d'eau douce est un résultat attendu [37, 38], l'observation de séquences associées à *Rhodothermus* phage RM378 l'est beaucoup moins, ce phage infectant une bactérie marine, *a priori* absente des milieux d'eau douce tempérés [39]. Pour aller au-delà de l'affiliation à partir d'un seul gène, il est possible de réaliser des graphiques de recrutement, associant les séquences des viromes à un génome de référence et permettant d'estimer le nombre de gènes du génome effectivement retrouvé dans les virus de l'écosystème étudié. Dans le cas du génome de *Rhodothermus* phage RM378,

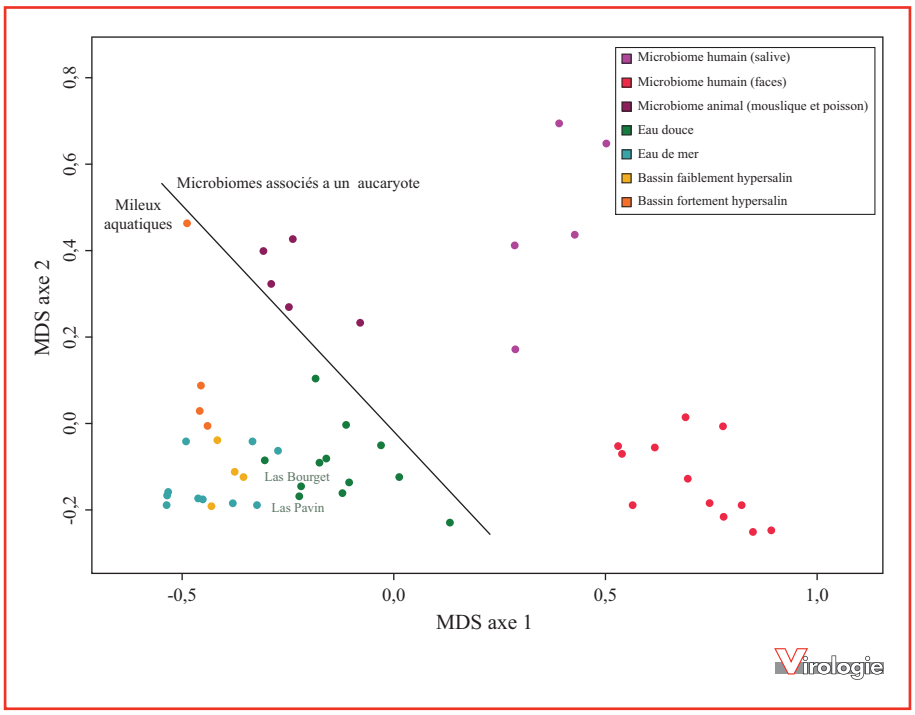


Figure 6. Distribution des différents viromes à partir d'une comparaison globale des séquences. Chaque virome est représenté par un point, coloré en fonction du type d'échantillon. La distance entre les points reflète la similarité entre les séquences des viromes considérés : ainsi, plus deux viromes se ressembleront en termes de séquence, plus ils seront proches sur le graphique.
Figure issue de Roux *et al.*, 2012 [22].

les similarités entre ce génome et les séquences de viromes lacustres sont clairement limitées à un nombre très restreint de gènes *(figure 8)*. De plus, les pourcentages d'identité entre le génome et les séquences du virome sont plutôt faibles (autour de 70 % au maximum). Les gènes retrouvés sont des gènes bien caractérisés (en rouge sur le graphique, en opposition aux gènes dont la fonction est inconnue, en bleu), et codant pour les fonctions nécessaires au développement du phage (notamment la réplication du génome et l'assemblage de la capside). Ainsi, les virus lacustres ayant un gène *TerL* dont la plus proche référence est le phage RM378 n'ont pas une composition très similaire à ce phage. Cet exemple illustre l'absence de référence proche pour la plupart des virus environnementaux et également la forte plasticité des génomes viraux en général et des génomes de phages en particulier.

Assemblage des viromes et création de génomes complets

Si les études métagénomiques apportent des informations inédites quant à la diversité virale dans l'environnement, la portée de telles analyses effectuées à partir des séquences brutes des viromes reste limitée puisque ces séquences correspondent au mieux à un gène complet. L'assemblage de ces mêmes séquences en contigs permet de dépasser cette limite et d'analyser de véritables fragments génomiques, jusqu'à plusieurs dizaines de milliers de nucléotides, voire des génomes complets. Le résultat de l'assemblage est très variable, puisqu'il dépend de la profondeur de séquençage, mais aussi de la richesse du métagénome étudié et de la longueur des génomes de départ.

La famille des *Microviridae* constitue un bon exemple du type de génome complet qu'il est possible d'assembler [40]. Il s'agit en effet de petits bactériophages, dont le génome est formé d'un fragment circulaire d'ADN simple brin d'environ 5 000 pb. Jusqu'à aujourd'hui, ces bactériophages ont principalement été décrits à partir de culture sur *E. coli*, ainsi que lors d'étude de bactéries pathogènes de type *Chlamydia* ou parasites comme *Bdellovibrio bacteriovorus* [41, 42]. Dix-huit génomes complets de *Microviridae* ont été obtenus par ces approches d'isolement et de culture, et six ont été détectés sous forme de prophages dans des génomes bactériens. En réalisant l'assemblage des séquences d'un ensemble de viromes précédemment décrits et publiés, il a été possible de générer 81 nouveaux génomes

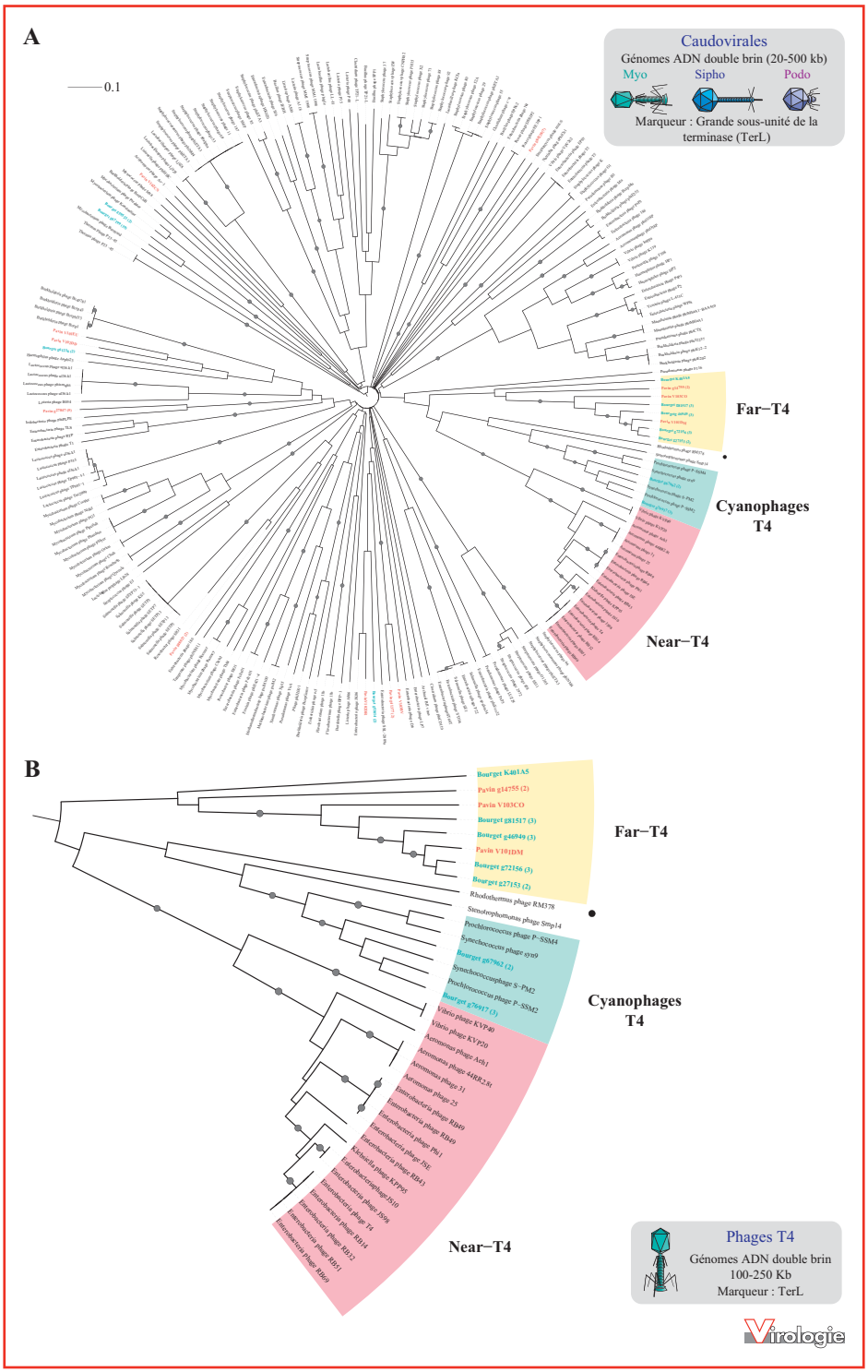


Figure 7. A) Arbre phylogénétique basé sur l'alignement des gènes codant pour la grande sous-unité de la terminase, gène commun à l'ensemble des Caudovirales (bactériophages à queue). Cet arbre permet d'associer des séquences de références (en noir) et des séquences issues des viromes (en rouge et bleu) ; **B)** zoom sur le sous-arbre comprenant la famille des phages T4. Les trois sous-groupes connus (near-T4, cyano-T4, Far-T4) sont indiqués sur l'arbre.
Figure issue de Roux *et al.*, 2012 [22] (avec l'aimable autorisation e M. Sime-Ngando, Society for Applied Microbiology and Blackwell Publishing Ltd, *Environmental Microbiology* 2013 ; 13 : 1956-72.).

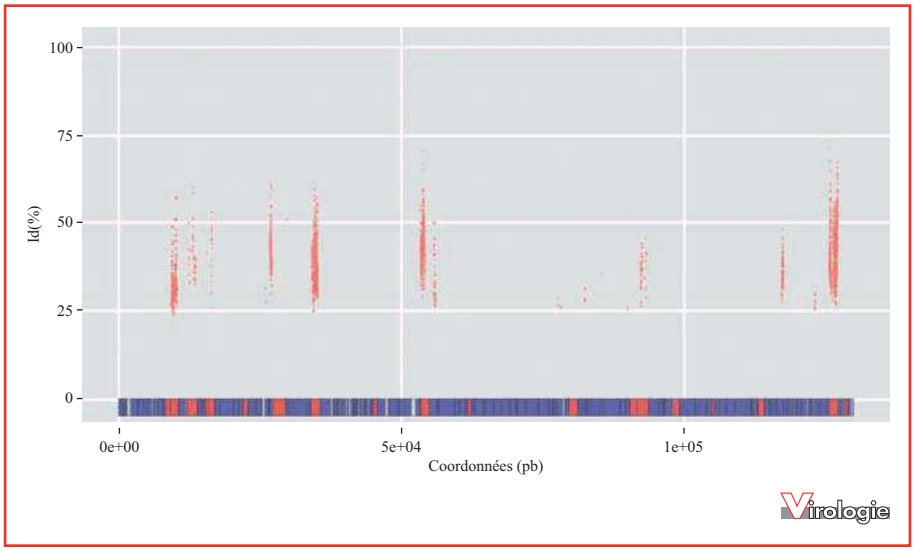


Figure 8. Graphique de recrutement des séquences du virome du lac du Bourget sur le génome de référence du groupe des Far-T4 (*Rhodothermus* phage RM378). La carte de génome est indiquée sur l'axe des abscisses, les gènes étant colorés en rouge quand leur fonction est connue, en bleu dans le cas contraire. Chaque séquence du virome similaire au génome est indiquée par un point, positionné en fonction de la zone de similarité (en abscisse) et du pourcentage d'identité entre les séquences du virome et le génomes (en ordonnée).

de *Microviridae*, dont 15 à partir des viromes des lacs Pavin et Bourget. Ces nouveaux génomes ont apporté de nouvelles connaissances au niveau de la structure, de la diversité et des modes d'évolution de cette famille.

Dans un premier temps, la phylogénie des *Microviridae* déjà connus et ceux assemblés à partir des viromes *(figure 9)* a permis de détecter une nouvelle sous-famille : les *Pichovirinae*. Cet arbre semble aussi témoigner d'une spécificité du sous-groupe des *Alpavirinae* pour le microbiome intestinal humain, tandis que les membres du groupe des *Gokushovirinae* sont retrouvés dans tous les types d'échantillons.

Ces génomes complets permettent également de constater une forte conservation des gènes et de l'ordre de ceux-ci le long du génome pour cette famille. L'existence d'un *core genome*, formé par trois gènes principaux (dont un impliqué dans la réplication du génome et deux dans la formation de la capside virale) a été confirmée. L'existence de transfert de gènes entre les *Microviridae* et leur hôte a aussi pu être mise en évidence grâce à la détection d'un gène codant pour une peptidase vraisemblablement d'origine bactérienne dans plusieurs des nouveaux génomes assemblés.

Enfin, les nouvelles séquences obtenues ont complété les données précédemment acquises quant à la structure des capsides de *Microviridae*. En particulier, l'évolution des protéines de capsides des *Microviridae* associés au microbiome humain semble spécifique, avec une accumulation de courtes insertions, pouvant aboutir à la formation de reliefs variables à la surface de la capside [40].

Conclusion

Grâce à la diminution des coûts de séquençage, le nombre de métagénomes viraux ne cesse d'augmenter. Ces données participent à une meilleure compréhension de la diversité des virus sur terre, des facteurs de régulation des communautés virales, ou encore de l'organisation et la plasticité des génomes viraux. De plus, les techniques de séquençage évoluent très rapidement, avec une augmentation quasi exponentielle du nombre de bases séquencées pour chaque échantillon. Ainsi certaines études utilisant les techniques de séquençage les plus récentes comme le séquenceur Illumina Hi-Seq 2000 ont pu reconstituer des génomes allant jusqu'à 76 Kb [43], ou étudier la conservation de cassettes de gènes au sein des génomes de bactériophages du tractus digestif humain [44]. À l'heure actuelle, de telles analyses ne sont possibles que pour les milieux dont la richesse génétique est la plus faible, mais à n'en pas douter ces méthodes seront applicables dans l'avenir aux communautés virales plus complexes comme celles des milieux océaniques.

Le champ d'application de la métagénomique virale va également en s'agrandissant, notamment en direction de recherches cliniques [45]. Ainsi, dès 2008, une étude de métagénomique virale a mis en évidence un nouveau type d'arénavirus transmis lors de transplantation d'organes [46]. Depuis, plusieurs autres associations entre pathologies et virus ont pu être révélées par approches métagénomiques

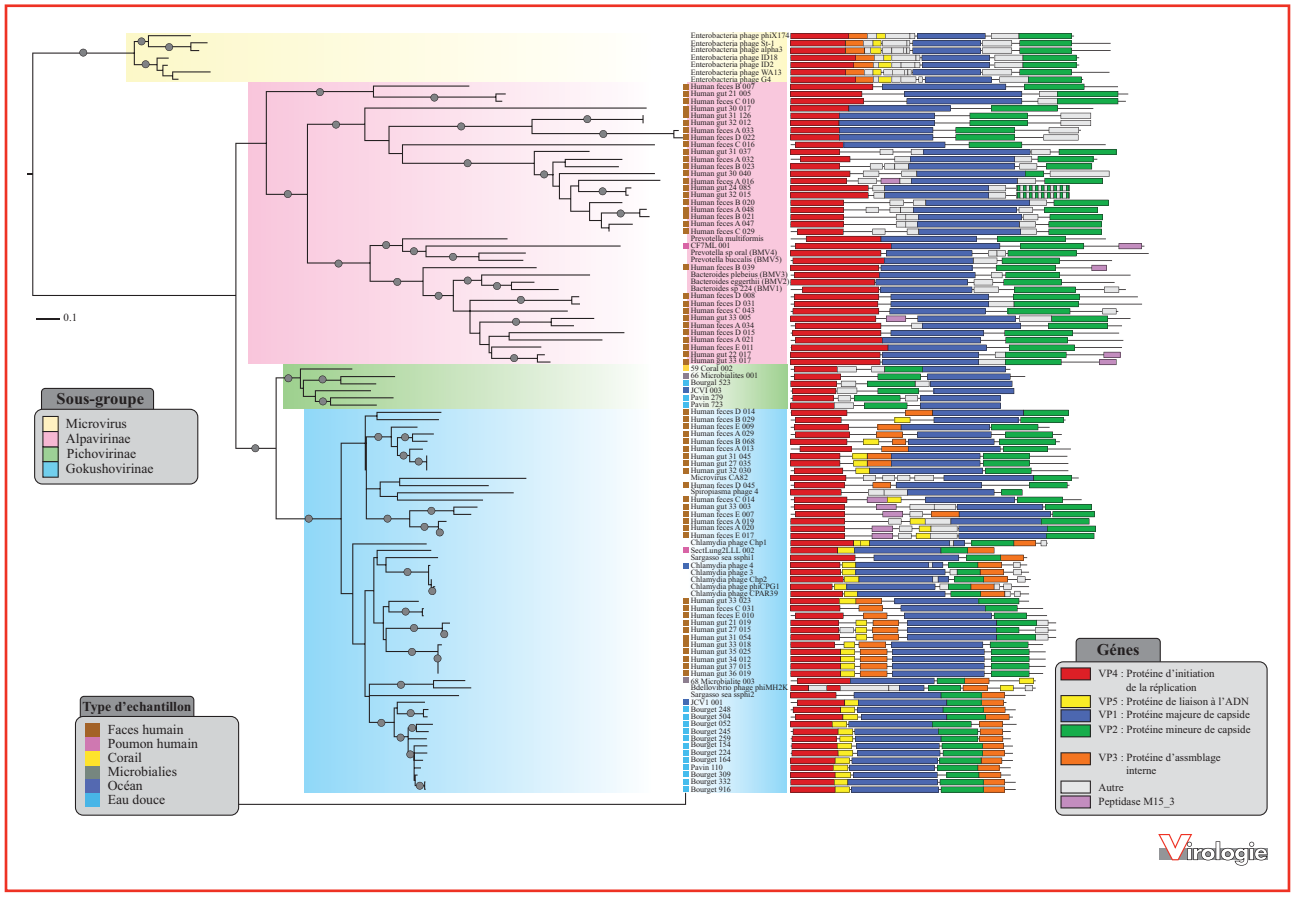


Figure 9. Arbre phylogénétique basé sur la protéine majeure de capside des génomes de *Microviridae* (petits virus à ADN simple brin et génome circulaire). Cet arbre comprend les génomes de *Microviridae* assemblés à partir de métagénomes (identifiés par un carré coloré en fonction du type d'échantillon d'origine). Les sous-groupes internes à la famille des *Microviridae* sont indiqués en couleur au sein de l'arbre. Les cartes des génomes de chaque virus sont représentées à droite de l'arbre.
Figure issue de Roux *et al.*, 2012 [40].

(voir par exemple [47-49]). Toutes ces approches sont encore exploratoires et en phase de développement, nécessitant une harmonisation et une réflexion globale tant au niveau des méthodes de préparation des échantillons que dans les analyses bio-informatiques post-séquençage, mais elles constituent néanmoins une base intéressante pour le développement de véritables protocoles complets et standardisés et, ainsi, une possible utilisation en « routine » de ces outils.

Enfin, si ces approches métagénomiques semblent prometteuses pour l'analyse des communautés virales, elles ont aussi mis en évidence la complexité de ces communautés et le manque actuel d'informations sur certains pans entiers de la virosphère. Certaines découvertes comme l'existence d'un génome apparemment issu de la recombinaison d'un virus ADN et d'un virus ARN [50] ont ainsi entraîné une remise en question des principes et théories quant à la séparation ancestrale entre les virus sur la base de la nature de leur génome, et plus généralement de la classification taxonomique des virus. Ainsi, l'analyse des viromes apporte de nouvelles questions quant à leur diversité, leur histoire évolutive et leur place dans la biosphère.

Liens d'intérêts : les auteurs déclarent n'avoir aucun lien d'intérêt en rapport avec l'article.

Références

1. Suttle CA. Viruses in the sea. *Nature* 2005 ; 437 : 356-61.

2. Forterre P, Soler N, Krupovic M, Marguet E, Ackermann HW. Fake virus particles generated by fluorescence microscopy. *Trends Microbiol* 2013 ; 21 : 1-5.

3. Sime-Ngando T, Lucas S, Robin A, *et al*. Diversity of virus-host systems in hypersaline Lake Retba, Senegal. *Environ Microbiol* 2010 ; 13 : 1956-72.

4. Abrescia NGA, Bamford DH, Grimes JM, Stuart DI. Structure unifies the viral universe. *Annu Rev Biochem* 2012 ; 81 : 795-822.

5. Hatfull GF, Jacobs-Sera D, Lawrence JG, *et al*. Comparative genomic analysis of 60 mycobacteriophage genomes: genome clustering, gene acquisition, and gene size. *J Mol Biol* 2010 ; 397 : 119-43.

6. Sencilo A, Paulin L, Kellner S, Helm M, Roine E. Related haloarchaeal pleomorphic viruses contain different genome types. *Nucleic Acids Res* 2012 ; 40 : 5523-34.

7. Yin Y, Fischer D. Identification and investigation of ORFans in the viral world. *BMC Genomics* 2008 ; 9 : 24.

8. Torrella F, Morita R. Evidence by electron micrographs for a high incidence of bacteriophage particles in the waters of Yaquina Bay, Oregon: ecological and taxonomical implications. *Appl Environ Microbiol* 1979 ; 37 : 774-8.

9. Bergh O, Børsheim KY, Bratbak G, Heldal M. High abundance of viruses found in aquatic environments. *Nature* 1989 ; 340 : 467-8.

10. Sandaa R, Short SM, Schroeder DC. Fingerprinting aquatic virus communities. In : Wilhem SW, Weinbauer MG, eds. *Manual of aquatic viral ecology*. Waco (Tx) : ASLO, 2010, p. 9-18.

11. Wilhelm SW, Carberry MJ, Eldridge ML, Poorvin L, Saxton MA, Doblin MA. Marine and freshwater cyanophages in a Laurentian Great Lake: evidence from infectivity assays and molecular analyses of *g20* genes. *Appl Environ Microbiol* 2006 ; 72 : 4957-63.

12. Breitbart M, Salamon P, Andresen B, *et al*. Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci U S A* 2002 ; 99 : 14250-5.

13. Breitbart M, Hewson I, Felts B, *et al*. Metagenomic analyses of an uncultured viral community from human feces. *J Bacteriol* 2003 ; 185 : 6220-3.

14. Angly FE, Felts B, Breitbart M, *et al*. The marine viromes of four oceanic regions. *PLoS Biol* 2006 ; 4 : e368.

15. Santos F, Yarza P, Parro V, Briones C, Antón J. The metavirome of a hypersaline environment. *Environ Microbiol* 2010 ; 12 : 2965-76.

16. Li L, Shan T, Wang C, *et al*. The fecal viral flora of California sea lions. *J Virol* 2011 ; 85 : 9909-17.

17. Kim MS, Park EJ, Roh SW, Bae JW. Diversity and abundance of single-stranded DNA viruses in human feces. *Appl Environ Microbiol* 2011 ; 77 : 8062-70.

18. Reyes A, Haynes M, Hanson N, *et al*. Viruses in the faecal microbiota of monozygotic twins and their mothers. *Nature* 2010 ; 466 : 334-8.

19. Dinsdale EA, Edwards RA, Hall D, *et al*. Functional metagenomic profiling of nine biomes. *Nature* 2008 ; 452 : 629-32.

20. Willner D, Thurber RV, Rohwer F. Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol* 2009 ; 11 : 1752-6.

21. Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. Are we missing half of the viruses in the ocean? *ISME J* 2013 ; 7 : 1-8.

22. Roux S, Enault F, Robin A, *et al*. Assessing the diversity and specificity of two freshwater viral communities through metagenomics. *PLoS ONE* 2012 ; 7 : e33641.

23. Roux S, Faubladier M, Mahul A, *et al*. Metavir: a web server dedicated to virome analysis. *Bioinformatics* 2011 ; 27 : 3074-5.

24. Vega Thurber R, Haynes M, Breitbart M, Wegley L, Rohwer F. Laboratory procedures to generate viral metagenomes. *Nat Protoc* 2009 ; 4 : 470-83.

25. Wommack KE, Sime-Ngando T, Winget DM, Jamindar S, Helton RR. Filtration-based methods for the collection of viral concentrates from large water samples. In : Wilhelm SW, Weinbauer MG, Suttle CA, eds. *Manual of aquatic viral ecology*. ASLO : Waco (Tx), 2010, p. 110-7.

26. John SG, Mendez CB, Deng L, *et al*. A simple and efficient method for concentration of ocean viruses by chemical flocculation. *Environ Microbiol Rep* 2011 ; 3 : 195-202.

27. Colombet J, Robin A, Lavie L, Bettarel Y, Cauchie HM, Sime-Ngando T. Virioplankton 'pegylation' : use of PEG (polyethylene glycol) to concentrate and purify viruses in pelagic ecosystems. *J Microbiol Methods* 2007 ; 71 : 212-9.

28. Lawrence J, Steward GF. Purification of viruses by centrifugation. In : Wilhelm SW, Weinbauer MG, Suttle CA, eds. *Manual of aquatic viral ecology*. ASLO : ASLO, 2010, p. 166-81.

29. Altschul SF, Gish W, Miller W, *et al*. Basic local alignment search tool. *J Mol Biol* 1990 ; 215 : 403-10.

30. Culley AI, Lang AS, Suttle CA. Metagenomic analysis of coastal RNA virus communities. *Science* 2006 ; 312 : 1795-8.

31. Whon TW, Kim MS, Roh SW, Shin NR, Lee HW, Bae JW. Metagenomic characterization of airborne viral DNA diversity in the near-surface atmosphere. *J Virol* 2012 ; 86 : 8221-331.

32. Minot S, Sinha R, Chen J, *et al*. The human gut virome: inter-individual variation and dynamic response to diet. *Genome Res* 2011 ; 21 : 1616-25.

33. Angly F, Rodriguez-Brito B, Bangor D, *et al*. PHACCS, an online tool for estimating the structure and diversity of uncultured viral communities using metagenomic information. *BMC Bioinformatics* 2005 ; 6 : 41.

34. Breitbart M, Rohwer F. Here a virus, there a virus, everywhere the same virus? *Trends Microbiol* 2005 ; 13 : 278-84.

35. Vega Thurber R. Current insights into phage biodiversity and biogeography. *Curr Opin Microbiol* 2009 ; 12 : 582-7.

36. Comeau AM, Krisch HM. The capsid of the T4 phage superfamily: the evolution, diversity, and structure of some of the most prevalent proteins in the biosphere. *Mol Biol Evol* 2008 ; 25 : 1321-32.

37. Dorigo U, Jacquet S, Humbert JF. Cyanophage diversity, inferred from *g20* gene analyses, in the largest natural lake in France, Lake Bourget. *Appl Environ Microbiol* 2004 ; 70 : 1017-22.

38. Chénard C, Suttle CA. Phylogenetic diversity of sequences of cyanophage photosynthetic gene *psbA* in marine and freshwaters. *Appl Environ Microbiol* 2008 ; 74 : 5317-24.

39. Blondal T, Hjorleifsdottir S, Aevarsson A, *et al*. Characterization of a 5′-polynucleotide kinase/3′-phosphatase from bacteriophage RM378. *J Biol Chem* 2005 ; 280 : 5188-94.

40. Roux S, Krupovic M, Poulet A, Debroas D, Enault F. Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads. *PLoS ONE* 2012 ; 7 : e40418.

41. Brentlinger KL, Hafenstein S, Novak CR, *et al*. *Microviridae*: a family divided: isolation, characterization, and genome sequence of PhiMH2K, a bacteriophage of the obligate intracellular parasitic bacterium. *J Bacteriol* 2002 ; 184 : 1089-94.

42. Cherwa J, Fane BA. *Microviridae*: microviruses and gokushoviruses. In : *eLS*. Chichester : John Wiley, 2011.

43. Emerson JB, Thomas BC, Andrade K, Allen EE, Heidelberg KB, Banfield JF. Metagenomic assembly reveals dynamic viral populations in hypersaline systems. *Appl Environ Microbiol* 2012 ; 78 : 6309-20.

44. Minot S, Wu GD, Lewis JD, Bushman FD. Conservation of gene cassettes among diverse viruses of the human gut. *PLoS ONE* 2012 ; 7 : e42342.

45. Fancello L, Raoult D, Desnues C. Computational tools for viral metagenomics and their application in clinical research. *Virology* 2012 ; 434 : 162-74.

46. Palacios G, Druce J, Du L, *et al*. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med* 2008 ; 358 : 991-8.

47. Victoria JG, Kapoor A, Li L, *et al*. Metagenomic analyses of viruses in stool samples from children with acute flaccid paralysis. *J Virol* 2009 ; 83 : 4642-51.

48. Sullivan PF, Allander T, Lysholm F, *et al*. An unbiased metagenomic search for infectious agents using monozygotic twins discordant for chronic fatigue. *BMC Microbiol* 2011 ; 11 : 2.

49. McMullan LK, Frace M, Sammons SA, *et al*. Using next generation sequencing to identify yellow fever virus in Uganda. *Virology* 2012 ; 422 : 1-5.

50. Diemer GS, Stedman KM. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol Dir* 2012 ; 7 : 13.

Annexe A.3 : Article

Methodological biases in coral viromics

Elisha M. Wood-Charlson[a], Karen D. Weynberg[a], Curtis A. Suttle[b], Simon Roux[c,d], Madeleine J. H. van Oppen[a,*]

[a] Australian Institute of Marine Science, PMB 3, Townsville MC, Queensland 4810, Australia
[b] Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada
[c]Laboratoire Micro-organismes : Genome and Environment, UMR CNRS 6023, Université Blaise Pascal, Clermont Université, Aubière, France
[d]CNRS, UMR 6023, LMGE, F-63177 Aubière

Soumis à **Plos ONE**

Elisha M. Wood-Charlson[a], Karen D. Weynberg[a], Curtis A. Suttle[b], Simon Roux[c,d], Madeleine J. H. van Oppen[a,*]

[a] Australian Institute of Marine Science, PMB 3, Townsville MC, Queensland 4810, Australia
[b] Department of Earth and Ocean Sciences, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada
[c] Laboratoire Micro-organismes : Genome and Environment, UMR CNRS 6023, Université Blaise Pascal, Clermont Université, Aubière, France
[d] CNRS, UMR 6023, LMGE, F-63177 Aubière

keywords : coral, disease, metagenome, virome, virus

Abstract

Our knowledge of the viral communities associated with the coral holobiont (the coral host animal and all of its eukaryotic and prokaryotic symbionts) is still in its infancy. Much of the pre-sequencing (sample preparation) and post-sequencing (bioinformatics) methodologies employed in coral virome studies originate from work with viruses, typically phages, in simpler systems, such as cultured lysates or seawater samples. The coral holobiont has many layers of potential viral hosts, incorporating all domains of cellular life, and this complexity will require the modification of current methods and understanding of their limitations, before we can begin to unveil viral diversity and function within the coral holobiont. Our meta-analysis uses available coral-holobiont sequence data to identify trends and potential complications inherent in the current methods used for studying viruses associated with corals. We show that various methods previously employed for isolation of viruses and their nucleic acids show clear biases to certain viral groups. Direct comparisons between studies are therefore not always justified. Inferences about certain groups being abundant members of coral-associated viral communities may be methodological artefacts and require confirmation. Further, the small number of viral reference genomes available in public sequence databases, as well as the short sequence read lengths typical for some next generation sequencing methods, are potential causes of erroneous virus identification. While metagenomic approaches are invaluable in progressing our understanding of coral-associated viruses, caution must be exercised in light of these pitfalls, and future studies should strive to combine modern and traditional virology methods.

Introduction

Scleractinian (stony) corals are the corner stone taxon of coral reefs. They are responsible for the three-dimensional structures that constitute coral reefs through the deposition of calcium carbonate during colony growth. This structural framework provides important habitats for a wide taxonomic diversity of coral reef organisms. Corals are also the main primary producers on coral reefs through photosynthesis of their single-celled, dinoflagellate endosymbionts of the genus *Symbiodinium*. These algal symbionts provide most of the coral's energy and promote calcification. Coral cover is decreasing world-wide due to a range of anthropogenic disturbances, including destructive fishing, sediment/nutrient/herbicide influx from terrestrial run-off, *Acanthaster planci* outbreaks (Adjeroud *et al.*, 2009; De'ath *et al.*, 2012; Sweatman *et al.*, 2011) and the impacts of greenhouse gas emissions such as increasing sea surface temperatures, ocean acidification and increased severity and frequency of cyclones (Carpenter *et al.*, 2008). There is a strong link between climate warming and the increased frequency and intensity of mass coral bleaching events, while a range of other factors are known to cause more localized bleaching (Berkelmans *et al.*, 2004; Hoegh-Guldberg, 1999; Wellington *et al.*, 2001; Winter *et al.*, 1998). A few studies suggest that viruses may be responsible for some instances of bleaching, but this remains to be verified (reviewed in van Oppen *et al.*, 2009).

The coral holobiont is a complex mini-ecosystem in which the coral animal acts as a host to a wide array of organisms from all domains of life, as well as the viruses that infect the eukaryotes and prokaryotes of the holobiont. Of all components in the coral holobiont, viruses are the least studied. In early publications, viral community descriptions were based on virus-like particle (VLP) morphologies examined through transmission electron microscopy (Davy *et al.*, 2006; Lohr *et al.*, 2007; Patten *et al.*, 2008; Wilson *et al.*, 2001; 2005). With the development of next generation sequencing methods over the past decade, an additional and higher-resolution metagenomics approach to viral community identification is now available (Bexfield and Kellam, 2011; Edwards and

Rohwer, 2005; Kristensen *et al.*, 2010; Rosario and Breitbart, 2011). However, there are inherent challenges to the purification of viruses and viral genomes from coral tissue. Removal of the tissue from the surface of the coral skeleton results in a thick mucus-like substance that is resistant to filtration/concentration methods commonly used in marine viral ecology. Instead, the sample is often subjected to chemical or mechanical disruption (Hick *et al.*, 2010; Marhaver *et al.*, 2008; Vega Thurber *et al.*, 2008) prior to loading onto a CsCl gradient for purification by centrifugation. The small size of virus particles and their genomes, combined with our inability to concentrate virus particles and the losses incurred during purification, results in low amounts of nucleic acids. Typically, whole genome amplification (WGA) is used to obtain enough material for sequencing; however, some methods, such as multiple displacement amplification (MDA), have been shown to introduce quantitative biases during replication (Angly *et al.*, 2006; Duhaime *et al.*, 2012; Yildiz and Visick, 2009). Moreover, the diverse genome chemistry of viruses (ssRNA, dsRNA, ssDNA, or dsDNA) complicates efficient and quantitative nucleic acid isolation from mixed communities (Andrews-Pfannkoch *et al.*, 2010). In addition to pre-processing samples for sequencing, marine viral metagenomes require careful post-sequencing processing. The majority of marine viruses lack representation in sequence databases; hence, many of the sequences remain unidentified (Angly *et al.*, 2006; Breitbart *et al.*, 2002; Breitbart *et al.*, 2007; Wegley *et al.*, 2007; Williamson *et al.*, 2008). Further, cellular sequences are often present in viral metagenomes, which may be the result of host contamination, horizontal gene transfer (HGT) events between viral genomes and their host, or gene transfer agents (GTAs) (Canchaya *et al.*, 2003; Jiang *et al.*, 1998; Koonin and Dolja, 2012; Lang *et al.*, 2012; Liu *et al.*, 2011; Monier *et al.*, 2009).

In spite of these caveats, metagenome and transcriptome sequence data provide qualitative information about the various members of the coral holobiont, and sequence data sets are available for several coral species. Regardless of the focus of these studies, be it the coral host tissue, the algal symbiont, or the prokaryotic microbes, it is very difficult to completely separate these target organisms from the associated viral community. Therefore, it is likely that the raw sequence data sets targeting a non-viral fraction will still contain viral sequences. Using this repository of sequence data, we examine how methodology, from the purification of viral particles to downstream sequence annotation and community characterization, can drastically influence conclusions about the composition and function of viral assemblages associated with the coral holobiont.

Materials and Methods

Sequence data and processing

Coral, prokaryotic and viral metagenome and meta-transcriptome data sets were downloaded from publically-available sequence archives of individual research labs and online resources including Metavir (Roux *et al.*, 2011), MG-RAST (Meyer *et al.*, 2008) and NCBI (Supplementary Table S1). In addition, unpublished partial RNA and DNA viromes from a single sample of *Pocillipora damicornis* were included for methodology comparisons.

Sequence data sets were uploaded to MG-RAST for basic characterization of the metagenomes, including the presence of ribosomal RNA genes and percent of sequences matching predicted proteins with known functions. Metagenomes were also interrogated using Metavir, an online resource designed to work specifically with viral metagenomic sequences (Roux *et al.*, 2011). This online tool provides taxonomic composition assessments, associated recruitment plots and phylogenetic trees, as well as virome to virome comparisons based on genetic richness of sequence similarity (details described under below). The Metavir taxonomic analyses presented in this study were based the 2012-07-10 NCBI viral refseq database.

Sequence analysis

Several versions of each sequence data set were created to allow for comparisons between studies that targeted cellular and acellular (i.e., viruses) compartments of the coral holobiont (Figure 1). Complete data sets were run through MG-RAST to estimate and remove rRNA contamination to obtain “clean

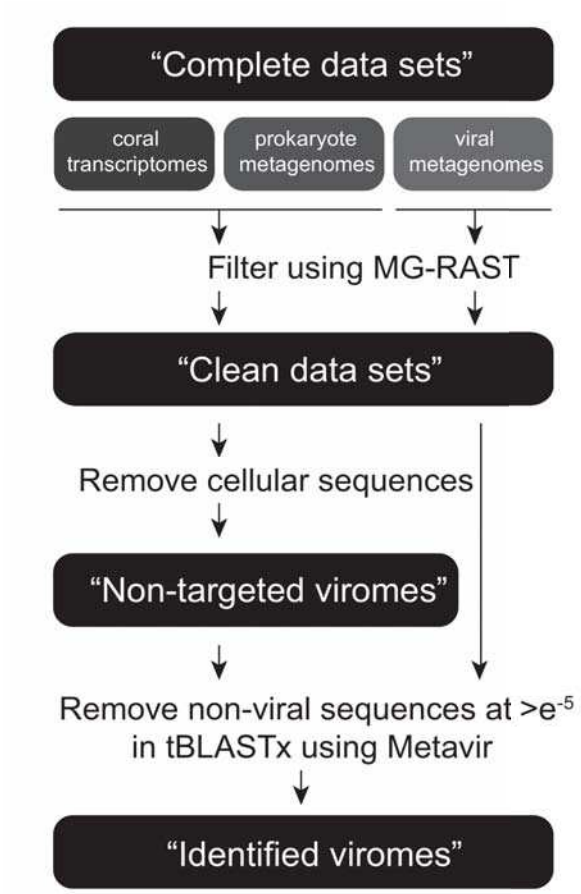


Figure 1. Schematic representation showing how the various sequence data sets were treated. Complete data sets: sequence data as were available in the public domain by various authors. These include metagenomes/transcriptomes, as well as viromes listed in Table 1 (publically available through MG-RAST under the project name "coral-associated viruses," unless otherwise noted). Clean data sets: data sets that contain sequences which passed quality control through MG-RAST and had rRNA sequences removed (publically available through Metavir under the project name "coral-associated viruses," unless otherwise noted). Non-targeted viromes: virome data sets obtained from metagenomes/transcriptomes targeted at cellular compartments of the coral holobiont and from which "cellular" sequences were removed (top hit "cellular", tBLASTx to NCBI nr database, e-value ≤10-3). Identified viromes: subset of sequences that were annotated by Metavir (e-value ≤10-5).

data sets", i.e. data sets without rRNA sequences (Table 1 indicates which data sets were added to or already available in MG-RAST). These clean data sets (Figure 1) were compared to the NCBI nr database (e-value $\leq 10^{-3}$) using tBLASTx. Any sequences that were annotated as "cellular" were removed, creating a subset of sequences representing "non-targeted viromes" – data sets that were not originally processed for viral metagenomic sequencing but have had sequences of cellular origin removed. Clean data sets, non-targeted viromes and viral metagenomes (Figure 1), were uploaded to Metavir for further comparison. Community structure was assessed at the level of genome organization (dsDNA, ssDNA, positive or negative ssRNA, dsRNA, retrovirus, satellite viruses and unclassified) and screened for the presence of known viral families (ICTV, 2012), using Metavir's BLAST comparison (e-value ≤ 10^{-5}) to NCBI viral refseq database. Metavir assigns a sequence to a viral taxonomic group when significant similarity is found, and then uses Genome-relative Abundance and Average Size (GAAS) to weight the taxonomic composition (Angly *et al.*, 2009) before the final assignment of a sequence to a viral taxonomic group. Viral metagenomes and identified viruses within non-targeted viromes were plotted as rarefaction curves (clustered at 98% similarity) to compare viral sequence diversity per unit sampling effort (Raes and Bork, 2008; Roux *et al.*, 2012). Data sets were also compared through clustering by dinucleotide abundances, which can be used to compare variation in genome signatures as a proxy for assessing similarity between meta-sequence data sets. Briefly, the dinucleotide frequency bias (relative frequency of each dinucleotide normalized by the frequency of each nucleotide in this dinucleotide) was calculated for each virome and then pairwise comparisons of viromes were computed ("city-block" metric distance normalized by the number of different nucleotide) (Willner *et al.*, 2009). The resulting distance matrix was then used in a meta-NonMetric Muldimensional Scaling (function metaMDS, package vegan, R software) to plot a 2D chart with each virome as a dot and distances between dots being as close as the distances in the matrix. Such comparison of k-mer nucleotide frequencies has been used to gather metagenomes according to their environments (Willner *et al.*, 2009), and more recently to classify contigs within a virome (Santos *et al.*, 2010) or compare closely related strains of *Caudovirales* and link them to putative hosts (Garcia-Heredia *et al.*, 2012)

Results and Discussion

In total, more than 2.4 million sequences from 11 coral species have been uploaded to Metavir (Roux *et al.*, 2011). We contributed 19

Species	Sequence type	Holobiont fraction	Complete data set # seq uploaded	MG-RAST # seq annotated	MG-RAST per cent rRNA	MG-RAST Accession	Cleaned data set # seq uploaded	Cleaned data set Avg. seq length	Metavir analysis # significant hit (e^{-5})	Metavir analysis # seq in identifed virome
Acropora hyacinthus[1]	Transcriptome – 2010	Coral	140 645	14.30%	2.40%	4492312.3[2]	140 645	<200bp	1.439%	2023
	Transcriptome – 2012	Coral	67 846	21.70%	0.10%	4492308.3[2]	67 846	<200bp	2.112%	1433
Acropora millepora[1]	Transcriptome – 2010	Coral	181 031	16.70%	0.90%	4492313.3[2]	181 031	<200bp	1.635%	2959
	Transcriptome – 2012	Coral	95 400	21.80%	0.10%	4492310.3[2]	95 400	<200bp	2.748%	2622
	Metagenome – Pre-bleach	Prokaryote	401 070	0.00%	11.40%	4445756.3[3]	401 070	>200bp	0.861%	3451
	Metagenome – Post-bleach	Prokaryote	403 686	2.10%	5.80%	4445755.3[3]	403 686	>200bp	1.187%	4790
Acropora palmata[1]	Transcriptome – 2011	Coral	88 020	40.80%	0.20%	4492315.3[2]	88 020	<200bp	5.884%	5179
Acropora tenuis[1]	Transcriptome – 2012	Coral	89 190	21.70%	0.10%	4492311.3[2]	89 190	<200bp	2.712%	2419
Diploria strigosa[1]	Metagenome – Healthy	Virus	1580	29.30%	0.00%	4487972.3[2]	1580	>200bp	22.910%	362
	Metagenome – Bleached	Virus	930	44.40%	0.00%	4487973.3[2]	930	>200bp	33.540%	312
Montastrea annularis[1]	Transcriptome	Coral	2 173	34.00%	1.70%	4492318.3[2]	2 173	>200bp	4.410%	96
Montastraea cavernosa[1]	Metagenome – Control	Virus	60 228	6.90%	2.90%	4455158.3[3]	56 218	varies	2.944%	1655
	Metagenome – Temperature	Virus	113 274	4.80%	11.50%	4455159.3[3]	104 166	varies	2.535%	2641
Montastrea faveolata[1]	Transcriptome	Coral	21 096	41.00%	4.10%	4492319.3[2]	21 096	>200bp	5.947%	1255
Porites astreoides[1]	Transcriptome – 2010	Coral	92 142	21.80%	2.50%	4492309.3[2]	92 142	<200bp	3.600%	3317
	Transcriptome – 2012	Coral	50 205	29.80%	0.10%	4492314.3[2]	50 205	<200bp	3.852%	1934
	Metagenome	Prokaryote	316 279	5.80%	3.90%	4440319.3[3]	295 437	<200bp	0.409%	1208
Porites compress[4]	Metagenome – Control	Virus	39 340	15.90%	2.90%	4440374.3[3]	39 191[4]	<200bp	0.805%	315
	Metagenome – DOC	Virus	35 680	4.50%	3.80%	4440370.3[3]	35 409[4]	<200bp	0.847%	300
	Metagenome – Nutrient	Virus	34 433	4.00%	4.10%	4440377.3[3]	34 139[4]	<200bp	0.967%	330
	Metagenome – pH	Virus	50 364	1.00%	5.80%	4440371.3[3]	49 949[4]	<200bp	0.828%	414
	Metagenome – Time 0	Virus	39 270	7.50%	2.10%	4440376.3[3]	39 113[4]	<200bp	0.659%	258
	Metagenome – Temperature	Virus	39 036	0.00%	7.60%	4440375.3[3]	38 482[4]	<200bp	0.254%	98
Pocillipora damicornis[1]	Metagenome – DNA	Virus	62 270	32.40%	2.40%	4492317.3[2]	62 270	>200bp	31.002%	19305
	Metagenome – RNA	Virus	80 787	54.60%	2.10%	4492316.3[2]	80 787	>200bp	12.179%	9839
Coral atolls: seawater[4]	Metagenome – Seawater	Virus								
Kingman Island	(remote)		94 915	6.50%	1.50%	4440036.3[3]	93 744[4]	<200bp	1.371%	1285
Palmyra Island	(remote)		358 983	2.20%	1.20%	4440040.3[3]	318 178[4]	<200bp	2.346%	7463
Tabueran Island	(inhabited)		380 355	2.20%	1.40%	4440280.3[3]	378 475[4]	<200bp	2.316%	8765
Kiritimati Island	(inhabited)		283 390	20.30%	2.40%	4440038.3[3]	279 882[4]	<200bp	0.904%	2531

[1] available in Metavir as "Coral-associated viruses" project
[2] available in MG-RAST as "Coral-associated viruses" project
[3] already available in MG-RAST
[4] already available in Metavir

Table 1. Summary statistics for the sequence data used in this study. Complete data sets were uploaded as provided, and the identified viral data sets represent a subset of those sequences that were annotated through Metavir (e-value ≤10-5). Information regarding the source of each data set is available in Supplementary Table 1.

transcriptomes/metagenomes to Metavir, which already contained six viromes from a stress experiment on the coral *Porites compressa* (Vega Thurber *et al.*, 2008) and four viromes from seawater collected on coral reefs in the Northern Line Islands (Dinsdale *et al.*, 2008) (Supplementary Table S1). As with previous marine virus sequencing studies (Breitbart *et al.*, 2007), the majority of sequences in these data sets (~97%) did not have a significant match (e-value $\leq 10^{-5}$) to a viral genome reference sequence (Table 1). However, viromes where the majority of reads were longer than 200 bp showed a larger proportion of significant hits (>12-33%), as was observed previously by (Wommack *et al.*, 2008).

Several versions of each sequence data set were created to compare among studies that targeted cellular and viral compartments of the coral holobiont (Figure 1). Complete data sets (available as a public project "Coral-associated viruses") were uploaded to or already available through MG-RAST (Meyer *et al.*, 2008). The summary statistics showed that all of the data sets – coral, prokaryote and viral – contained at least some rRNA genes, except for the linker-amplified viromes produced for healthy and bleached *Diploria strigosa* (Marhaver *et al.*, 2008, discussed later) (Table 1). Many of the data sets included in this analysis were not directed at sequencing the viral component of the holobiont, so cellular sequences were expected. However, almost all of the coral-associated viromes also contained rRNA genes (1.2-11.5%) indicating some level of contamination by cellular nucleic acids. Cellular contamination seems to be a pervasive problem (Kristensen *et al.*, 2010), despite attempts to remove cellular material by chloroform addition, separation of viral particles by density-gradients and centrifugation, filtration and/or size-fractionation and DNase/RNase treatment prior to release of the viral nucleic acids from their capsids (Correa *et al.*, 2013; Dinsdale *et al.*, 2008; Vega Thurber *et al.*, 2008).

Community composition is influenced by laboratory and bioinformatic methods

Clean data sets and non-targeted virome data sets (Figure 1) were submitted to Metavir (Roux *et al.*, 2011). We chose a moderately conservative BLAST expect value (e-value) of $\leq 10^{-5}$ because several data sets (available in Metavir under *P. compressa* project) differed drastically in community composition when e-values between $\leq 10^{-3}$ and $\leq 10^{-5}$ were used, but less so between $\leq 10^{-5}$ and $\leq 10^{-7}$. The discrepancies seen in the "low-similarity" BLAST hits (e-value $>10^{-3}$) likely result from

the lack of references genomes for marine viruses, leading to low levels of similarity and unstable taxonomic affiliation. To test whether cellular sequences impacted viral community identification through Metavir, taxonomic composition of clean data sets, which still contained cellular sequences, and the non-targeted viromes (Figure 1), were compared. Minimal to no differences in taxonomic composition were observed (data not show), suggesting that the identifiable viral communities were not obscured by the presence of cellular sequences.

Except for viromes with longer than 200 bp sequence reads, the proportion of significant hits (e-value $\leq 10^{-5}$) to a "known" viral sequence was typically less than 3% (Table 1) (Angly *et al.*, 2006; Breitbart *et al.*, 2002; Breitbart *et al.*, 2007; Wegley *et al.*, 2007; Williamson *et al.*, 2008). Hence, conclusions about viral communities associated with the coral holobiont have been drawn from a very small number of often short sequence reads, and these conclusions should be treated with caution until findings are confirmed. The one published exception was the *D. strigosa* data set (Marhaver *et al.*, 2008) with 22-33% of sequences giving a significant hit to a known viral sequence. Compared to the other coral viromes, this data

Type	Order	Family	Host	*A. hyacinthus* 2010	*A. hyacinthus* 2012	*A. millepora* 2010	*A. millepora* 2012	*A. palmate* 2011	*A. tenuis* 2012	*M. annularis*	*M. faveolata*	*P. asteroides* 2010	*P. asteroides* 2012	*P. asteroides* prokaryote	*A. millepora* (pre-bleach)	*A. millepora* (post-bleach)	*D. strigosa* (healthy)	*D. strigosa* (bleached)	*M. cavernosa* control	*M. cavernosa* temp	*P. compressa* control	*P. compressa* DOC	*P. compressa* nutrient	*P. compressa* pH	*P. compressa* t0	*P. compressa* temp	*P. damicornis* RNA	*P. damicornis* DNA	Kingman seawater	Palmyra seawater	Tubuean seawater	Kiritimati seawater
ds DNA I	Caudovirales	Myoviridae	B																													
		Podoviridae	B																													
		Siphoviridae	B																													
	Herpesvirales	Alloherpesviridae	V																													
		Herpesviridae	V																													
		Malacoherpesviridae	I																													
	Unassigned	Adenoviridae	V																													
		Ascoviridae	I																													
		Asfarviridae	V																													
		Baculoviridae	I																													
		Bicaudaviridae	Ar																													
		Corticoviridae	B																													
		Fuselloviridae	Ar																													
		Iridoviridae	I, V																													
		Lipothrixviridae	Ar																													
		Mimiviridae	Pr																													
		Nimaviridae	I																													
		Papillomaviridae	V																													
		Phycodnaviridae	AI																													
		Polydnaviridae	I																													
		Polyomavirus	V																													
		Poxviridae	I, V																													
		Rudiviridae	Ar																													
	Unclassified																															
ss DNA II	Unassigned	Circoviridae	V																													
		Geminivirdae	P																													
		Inoviridae	B																													
		Microviridae	B																													
		Nanoviridae	P																													
		Parvoviridae	I, V																													
		Unassigned																														
ds RNA III	Unassigned	**Chrysoviridae**	F																													
		Cystoviridae	B																													
		Partitiviridae	F, P, Pr																													
		Reoviridae	F, I, P, V																													
	Unclassified																															
ss RNA (+) IV	Nidovirales	**Coronaviridae**	V																													
	Picornavirales	**Picornaviridae**	V																													
		Secoviridae	P																													
	Tymovirales	**Tymoviridae**	F, P																													
	Unassigned	**Bromoviridae**	P																													
		Flaviviridae	I, V																													
		Narnaviridae	F																													
		Nodaviridae	I, V																													
		Potyviridae	P																													
		Togaviridae	*I, V*																													
		Tombusviridae	P																													
	Mononegavirales	Unassigned																														
ss RNA (-) V	Unassigned	Unassigned																														
		Bunyaviridae	I, P, V																													
ss RNA (+) RT VI	Unassigned	**Retroviridae**	V																													
dsDNA RT		**Caulimoviridae**	P																													
Unclassified ss RNA		Dinornavirus	Pr																													
Unclassified archael			Ar																													
Virophages																																
Satellite																																
				CORAL TRANSCRIPTOME										PROKARYOTE METAGENOME			VIROME													SEAWATER VIROME			

Table 2. The presence of viral families in each of the complete data sets used in this study. Bold family names identify those with a known marine isolate (ICTV, 2012; Steward et al., 2013). Shaded box indicates that viral family was present in the Metavir analysis (e-value ≤10-5). Information regarding the source of each data set is available in Supplementary Table 1 .

set was quite small, but it contained pre-screened high-quality sequences with the majority of sequences greater than 400 bp in length, a factor known to increase sequence identification (Wommack *et al.*, 2008). The relatively high proportion of viral hits in the preliminary *P. damicornis* viromes (12-31%), which consisted of sequences with an average read length of 680 bp, supports this.

Our qualitative comparison of the viral community associated with the coral-derived metagenomes and transcriptomes clearly demonstrates the impact of sample collection and processing in determining downstream results. The coral transcriptomes were dominated by reverse transcribing (RT) viruses (~52-75%, Supplementary Table S2), including *Retroviridae* (ssRNA) and *Caulimoviridae* (dsDNA), but were almost devoid of ssDNA viruses (Table 2). The coral-associated metagenomes (host, prokaryote, or virus) were dominated by dsDNA or ssDNA viruses. The relative dominance of dsDNA or ssDNA viruses in these samples must be viewed with caution as almost all of these metagenomes underwent whole genome amplification (WGA) using a standard GenomiPhi reaction prior to sequencing (Yokouchi *et al.*, 2006). The single non-WGA exception were the viromes for healthy and bleached *D. strigosa*, which were created using sheared genetic material ligated to linker sequences and randomly amplified using linker-specific primers (Marhaver *et al.*, 2008). This method is used to amplify dsDNA (Duhaime *et al.*, 2012; Henn *et al.*, 2010; Kim and Bae, 2011), so without conversion of other viral genotypes to dsDNA (Andrews-Pfannkoch *et al.*, 2010; Culley and Suttle, 2010), the rest of the viral community will likely be lost (reflected in Table 2). Although ssDNA viruses have been shown to be abundant in other marine viromes, and likely play an important role in the marine environment (Angly *et al.*, 2006; Liu *et al.*, 2011; Rosario *et al.*, 2012), their dominance in some of the coral-associated data sets may be an artefact of MDA, such as GenomiPhi (discussed in Angly *et al.*, 2006; Wegley *et al.*, 2007). It may be possible to smooth out some of the initial stochastic biases of MDA for ssDNA genomes. Dinsdale *et al*. (2008) performed 6-8 replicate GenomiPhi reactions for sample amplification and their results showed similar taxonomic patterns between sampling sites, without the over-representation of a single, dominant ssDNA representative (as seen in Supplementary Figure 1), suggesting the observed pattern may reflect the natural community more accurately than patterns derived from single GenomiPhi reactions (Supplementary Table 2). Other methods for discovering novel viruses are available, though several of the techniques, such as PCR using degenerate primers and microarrays, require *a priori* knowledge of the viral targets (reviewed in Bexfield and Kellam, 2011) and are not designed to identify viral communities. Linker-amplification techniques have recently become an alternative to MDA for amplifying nucleic acids prior to sequencing viral communities present in environmental samples (Ambrose and Clewley, 2006; Culley and Suttle, 2010; Djikeng *et al.*, 2008; Duhaime *et al.*, 2012).

The only group of viruses (based on the nature of their genome structure) that were present in all data sets were the dsDNA viruses, mostly from the order *Caudovirales* (Table 2). *Caudovirales* contains three families of tailed bacteriophages that likely infect members of the rich bacterial community associated with the coral mucus layer, gastric cavity and even some that are internal to the coral tissue (Agostini *et al.*, 2012; Sweet *et al.*, 2011). Re-analysis of just the published coral-associated viromes (Correa *et al.*, 2013; Dinsdale *et al.*, 2008; Marhaver *et al.*, 2008; Vega Thurber *et al.*, 2008) revealed that dsDNA and ssDNA bacteriophages had the most abundant reads with a significant match in the database (Supplementary Table S2), with the majority belonging to *Caudovirales* and *Microviridae*, respectively. Although bacteriophages were not addressed in several of the publications (Correa *et al.*, 2013; Vega Thurber *et al.*, 2008), our findings that bacteriophages dominate the viral community agree with the analyses by Dinsdale *et al.* (2008) and Marhaver *et al*. (2008), as well as other DNA-based aquatic viromes (Angly *et al.*, 2006; Breitbart *et al.*, 2002; Hewson *et al.*, 2012; Roux *et al.*, 2012; Steward and Preston, 2011). While amplification methods prior to sequencing may lead to quantitative biases in the viral community structure, a high proportion of bacteriophages is not surprising since corals

contain a diverse and prolific community of prokaryotes (Bourne and Munn, 2005; Ducklow and Mitchell, 1979; Kooperman *et al.*, 2007; Littman *et al.*, 2009; Raina *et al.*, 2009; Ritchie and Smith, 2004; Rohwer *et al.*, 2001; Rohwer *et al.*, 2002; Rosenberg *et al.*, 2007). Further, it should be noted that the laboratory methods used to create coral-associated viromes rely on isolation techniques developed for purifying a single cultured bacteriophage from a host cell lysate using CsCl-density gradients (Lawrence and Steward, 2010; Sambrook and Russell, 2001). In general, viruses are typically found within the 1.18-1.51 g/cm^3 density layer in a CsCl gradient (Lawrence and Steward, 2010; Vega Thurber *et al.*, 2009). However, published methodologies for examining coral-associated viral communities recommend targeting the 1.35-1.50 g/cm^3 CsCl density layer (Vega Thurber *et al.*, 2009), which is the density fraction typical for bacteriophages. Collection of only this density fraction should help reduce cellular contamination (cell buoyant density < 1.35 g/cm^3) (Lawrence and Steward, 2010), but many eukaryotic viruses also have buoyant densities <1.35 g/cm^3 (Vega Thurber *et al.*, 2009). So, while bacteriophages were expected in these viromes, they were also preferentially selected for using the current methodology.

In the majority of the coral-associated viromes in this study, a small number of reads were similar to viruses infecting eukaryotes, including sequences resembling members of the *Ascoviridae*, *Iridoviridae*, *Mimiviridae* and *Phycodnaviridae* families. *Iridoviridae* (1.26-1.60 g/cm^3), *Mimiviridae* (~1.36 g/cm^3) and *Phycodnaviridae* (~1.2-1.4 g/cm^3) have identified marine viral representatives and, given their densities, can occur within the typical 1.35-1.5 g/cm^3 fraction (ICTV, 2012; Jacobsen *et al.*, 1996; Steward *et al.*, 2013; Thurber and Correa, 2011). Although *Ascoviridae* evolved from the *Iridoviridae* family (Bigot *et al.*, 2009; Stasiak *et al.*, 2003), currently it contains only one recognized genus that infects lepidopterans, so the physical properties of a potential marine ascovirus-like counterpart are unknown. The published data from the *P. compressa* stress experiments focused on eukaryotic viruses by using a boutique sequence database that included reference genomes for eukaryotic viruses only (Vega Thurber *et al.*, 2008). Re-analysis of the *P. compressa* data sets through Metavir, which compares sequences to all viral reference genomes, showed that, of the total number of viral sequences with a hit to the databases (e-value $\leq 10^{-5}$, <1% of the number of sequences obtained), 48-92% matched to bacteriophages (Supplementary Figure S1), which were not discussed in the paper. A more recent study targeting the RNA virome of *Montastraea cavernosa* also used a eukaryotic virus boutique database to identify the viral community present, with the phage and archaeal virus communities to be presented elsewhere (Correa *et al.*, 2013). Again, analysis of these recent data sets through Metavir (e-value $\leq 10^{-5}$) against a database that includes all available viral reference sequences identified the majority of viruses in the published data set as bacteriophages (74-75%). Considering that most sequences in a virome cannot be identified because of the limited number of viral reference sequences, we would like to encourage any coral-associated viral metagenomic analyses to include the largest reference database possible, especially because the microbial community associated with corals is known to be diverse and important to the holobiont (Rosenberg *et al.*, 2007). In addition, as the size of the database decreases, a BLAST e-value becomes less reliable. This is because the expectation for finding a match (with equal similarity) between a metagenome sequence and a sequence in the database by chance is reduced in a small database compared to larger databases. Therefore, BLAST comparisons to restricted databases should use thresholds that are independent of the database size (i.e. BLAST bit-score), or if using database-dependent thresholds, such as e-values, they need to be adjusted appropriately.

As an additional complication, the underrepresentation of virus genomes in the public databases can cause sequence hits to genes of certain viruses that are not closely related to the virus from which the sequences originated. This is well-illustrated by the OtV-1 and OtV-2 genomes (viruses in the size range of 100-120 nm that infect the prasinophyte alga *Ostreococcus tauri*), which have at least three annotated genes that have closest significant

identity to genes found in the giant virus Mimivirus but are not described in any other virus to date (Weynberg *et al.*, 2009; 2011). A recent coral virome study showed several highly significant hits (65 hits in control and 129 hits in heat-treated data set) to genes in giant viruses of the family *Mimiviridae* (Correa *et al.*, 2013). However, this study included a 0.22 µm filtration step that most likely removed giant viruses (diameter ≥750 nm) prior to sequencing (Xiao *et al.*, 2005). This example shows that caution must be exercised when interpreting results from coral viromic approaches, especially until more marine and particularly coral-associated viral genomes (none available to date) have been sequenced. In addition, we encourage researchers to complement their coral metagenome sequencing efforts with other methods commonly used in marine viral ecology, such as flow cytometry and TEM, which can be used to verify abundance and identity of viral groups discovered through sequence analysis.

Are Herpesvirales *common viruses of the coral holobiont?*

With the exception of the *Herpesvirales* (1.22-1.28 g/cm^3), the lighter density dsDNA marine viruses are not well represented in the current virome data sets (Table 2) (Vega Thurber *et al.*, 2009). This may be an accurate reflection of the community present in the original samples, or it could be an artefact of isolation methods which have selected for more dense viruses. The prevalence of *Herpesvirales* in these viromes suggests that the former may be true; however, one other methodological caveat must be noted. Thus far, most of the coral-associated viromes available have used a chloroform step prior to the CsCl density

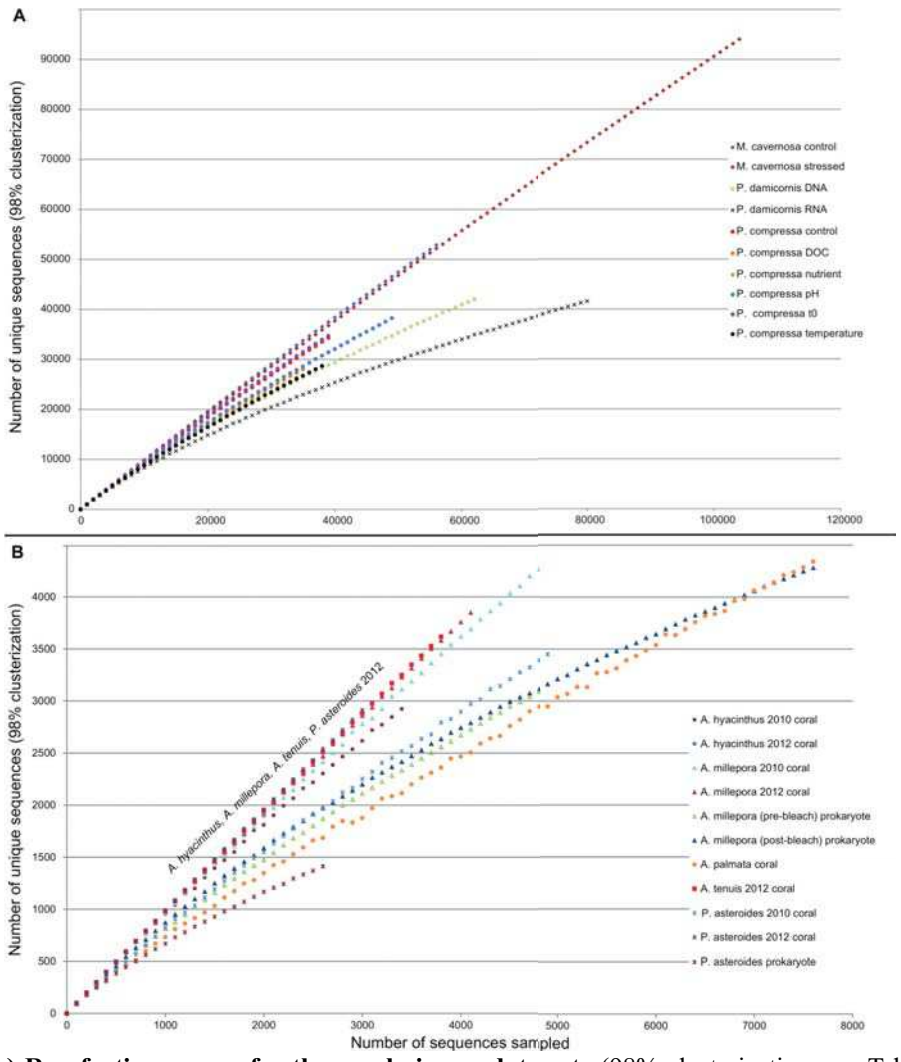


Figure 2. A) Rarefaction curves for the coral virome data sets (98% clusterization, see Table 1 for data set descriptions). **B) Rarefaction curves for the identified viromes within non-targeted viromes** (Metavir e-value ≤10-5, 98% clusterization).

gradient (Dinsdale *et al.*, 2008; Marhaver *et al.*, 2008; Vega Thurber *et al.*, 2008). This step removes much of the cellular contamination through membrane disruption, but chloroform can also affect the lipid-enclosed dsDNA eukaryotic viruses (Fields *et al.*, 2007). Unfortunately, little is known about how chloroform treatment changes the buoyant density of these viruses. There is, however, evidence to suggest that the *Herpesvirales* capsid (internal to the lipid membrane) is similar to the capsid of dsDNA bacteriophages and may even share common ancestry (Baker *et al.*, 2005; Homa and Brown, 1997). Therefore, after treatment with chloroform, marine *Herpesvirales*-like viruses may resemble dsDNA bacteriophages, which could explain their physical presence in the chloroformed 1.35-1.5 g/cm^3 CsCl density fraction (Marhaver *et al.*, 2008; Vega Thurber *et al.*, 2008). In a recent paper on the total RNA viromes isolated from one healthy and one temperature stressed *M. cavernosa* coral colony, chloroform treatment was omitted resulting in the viral fraction occurring closer to 1.2 g/cm^3 CsCl density layer (Correa *et al.*, 2013). Interestingly, the dsDNA viral community in this study included members of the lighter density dsDNA viruses, the majority of which were not from *Herpesvirales* (Table 2).

Regardless of the sample preparation method employed prior to sequencing, the bioinformatic perspective taken for sequence analysis can drastically affect data interpretation. For example, when compared to previously described eukaryotic viral communities in corals (Marhaver *et al.*, 2008; Vega Thurber *et al.*, 2008), our Metavir analyses showed a decreased emphasis on putative disease-causing viruses, such as members of *Herpesvirales*. Instead, the most common dsDNA eukaryotic viral family present in >80% of the data sets was the *Phycodnaviridae,* which agrees with the recent virome data from *M. cavernosa* (Correa *et al.*, 2013). Finding *Herpesvirales* in coral metagenomes is not surprising because the taxonomic Order contains several marine representatives that are known to cause disease in a variety of marine organisms, though mostly vertebrates (Hanson *et al.*, 2011; Maness *et al.*, 2011). If herpes-like viruses are infecting members of the coral holobiont, then the low

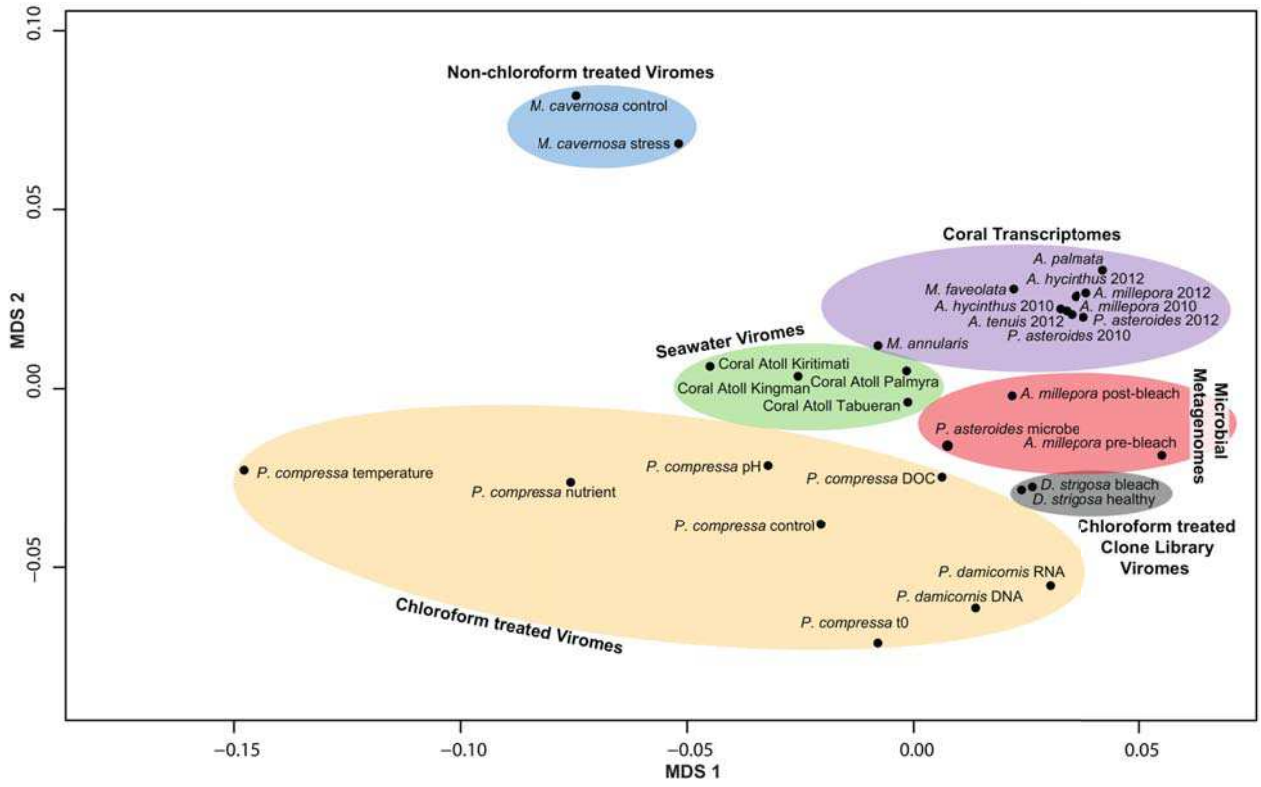


Figure 3. Dinucleotide clustering of the coral-associated virome and non-targeted virome data sets used in this study. Each oval represents a different method for sequence isolation: coral host transcriptome, prokaryotic metagenome, seawater viromes or coral viromes with and without chloroform treatment and sequenced by next-generation technology or from a DNA clone library.

abundance of *Herpesvirales* seen in our analyses may reflect the inability of current computational methods to match divergent coral-associated herpes-like viruses to the representative herpes genomes. Notexz in the coral data sets where *Herpesvirales* were documented (Correa *et al.*, 2013; Vega Thurber *et al.*, 2008) most of the sequences annotated to a herpes genome had low significance or were clustered in repeat regions of the genomes (highlighted in white, Supplementary Figure S2). Therefore, until a coral-associated herpes-like virus can be isolated, characterized and associated with a disease phenotype, the existence of coral-associated herpes-like viruses and their relationship to coral health remains unknown.

Preliminary patterns of diversity in coral-associated viral community

Rarefaction curves were used to estimate sequence richness for the viral metagenomes and identified viral communities from within the non-targeted viromes (Figure 1). None of the viromes appeared to achieve sampling saturation, but a few interesting trends can be seen from the rarefaction curves. The *P. damicornis* DNA and RNA libraries (Figure 1a) showed that the RNA viral library begins to reach sampling saturation more rapidly than the DNA viral library. Coral viromes may contain less RNA viral diversity, analogous to the relative diversity of RNA viruses to dsDNA viruses in the marine environment (reviewed in Steward *et al.*, 2013). In addition, after a natural coral bleaching event, colonies showed slightly higher sequence diversity than before bleaching (Figure 2b) (Littman *et al.*, 2011). Unfortunately, the identified viromes within this data set only represent 0.86-1.2% of the total sequences and do not include any novel viruses that may be present in the community (Table 1), so this slight signal may or may not reflect changes in the microbial community as a whole. The only other coral virome study that examined how natural stressors change the viral community was too small to compare via rarefaction curves (Marhaver *et al.*, 2008). More research is needed to determine how environmental stressors affect the coral-associated viral community and therefore the health of the coral holobiont. Do bleached corals typically show increased diversity of identifiable viruses and does the difference depend on sampling time or bleaching trigger (UV, temperature, salinity)? Which host might be the source for the increase in viral diversity (coral animal, algal symbiont, microbial community, or all three)? Is an observed change in diversity due to an outbreak from resident viruses, opportunistic viruses from the environment and/or latent viruses emerging from the host genomes? Does virus-induced coral bleaching occur in the natural environment? Small clues to these questions are starting to surface. A recent proteomic study in the coral *Stylophora pistillata* detected increased production of a viral replication protein in an algal symbiont enriched-fraction of the holobiont after exposure to metal depletion and elevated temperatures (Weston *et al.*, 2012), suggesting that this increase in viral replication may be at least partly due to *Symbiodinium*-associated viruses. This potential symbiont host-virus link will need to be confirmed from *Symbiodinium* cultures.

Genomes have been shown to possess signatures in oligonucleotide/dinucleotide abundances that reflect phylogenetic community composition in response to environmental selection (Kariin and Burge, 1995; Willner *et al.*, 2009). Metagenome clustering based on dinucleotide signatures has been used to explore the variation in microbial and viral metagenomes and was shown to perform the best for metagenomes with short reads (Willner *et al.*, 2009), similar to many of the data sets in this analysis. Dinucleotide clustering of the clean data sets in Metavir showed that the coral transcriptomes clustered closely together (Figure 3), likely reflecting their common methodology (discussed above) and the resulting viral composition (Supplementary Table S2). Although the non-targeted virome data sets contain cellular sequences, the dinucleotide cluster analysis for their identified viromes had a similar clustering pattern to the larger data sets (data not shown). The seawater viromes, processed with similar methodology as the coral viromes (i.e. use of chloroform) (Dinsdale *et al.*, 2008), formed a tight cluster away from the coral viromes, suggesting that coral-associated viruses are distinct from the overall marine viral

community. Finally, the recently published, non-chloroform treated coral viromes (Correa *et al.*, 2013) appear different from other data sets in this meta-analysis.

Our reanalysis of published data shows that any one method of viral nucleic acid isolation will only provide partial viromes. Therefore, even qualitative comparisons between studies in which different methods were used, may lead to inaccurate conclusions. We encourage the parallel isolation of DNA- and RNA-viromes for future studies of coral-associated viruses, as they represent distinct viral communities (Andrews-Pfannkoch *et al.*, 2010; Steward *et al.*, 2012), which was demonstrated in a recent paper on RNA and DNA viruses from a soft coral (Hewson *et al.*, 2012). In addition, a comparison of our own preliminary RNA- and DNA-viromes (methods are being optimized and will be reported elsewhere) isolated from a *P. damicornis* colony showed that the RNA virome contained viral families within the ds- and ssRNA viruses that were not uncovered in the DNA virome (Table 2).

Conclusions

Viruses play important roles as both pathogens and mutualists in many, if not all, forms of cellular life (Munn, 2006; Roossinck, 2011) and undoubtedly have important but currently unknown functions in the coral stress response, coral disease and the evolution of corals in response to climate change. Hence, in-depth studies of coral-associated viruses are urgently needed. The creation of marine viral metagenomes from complex biological associations, such as the coral holobiont, requires careful planning and testing of both laboratory and bioinformatic methodology. We have highlighted the current issues in coral viromics, while providing some suggestion on how these issues can be managed by judicious modification of current techniques. For example, sequencing both RNA and DNA viruses, understanding how laboratory methods may bias sequencing results and conducting analyses on the entire viral community, especially the bacteriophage population which appears to dominate coral viromes but to date has not been addressed in-depth. Others issues, like the low representation of marine viruses in sequence databases resulting in the majority of unknown viral sequences being discarded from post-sequencing analyses, high occurrences of horizontal gene transfer events between host and viral genomes and the likely presence of GTAs, will require a community effort to resolve. As part of the effort, recent online tools such as Metavir (Roux *et al.*, 2011) and VIROME (Wommack *et al.*, 2012) provide platforms that allow for ongoing comparisons between new and previously published virome data sets, as well as in-depth analyses based on up-to-date databases. However, as with any data set, the outcomes are only as good as the tools. Until our fundamental understanding of the coral viral community improves, our interpretation of the data should remain consistent with the chosen methodology and reflect the limitations of the study.

Acknowledgements

We thank Andy Muirhead for his assistance with the isolation of the *Pocillipora damicornis* viromes, Florent Angly for assistance with GAAS, Lionel Guidi for help with processing large data sets and Alexander Culley for comments on this manuscript. We would also like to thank two anonymous reviewers for their constructive suggestions. We acknowledge funding from the Australian Research Council (Future Fellowship #FT100100088 to MvO, Super Science Fellowship #FS110200034 to KW) and from the Australian Institute of Marine Science.

References

Adjeroud M, Michonneau F, Edmunds PJ, Chancerelle Y, de Loma TL, Penin L *et al*. (2009). Recurrent disturbances, recovery trajectories, and resilience of coral assemblages on a South Central Pacific reef. *Coral Reefs* **28:** 775-780.

Agostini S, Suzuki Y, Higuchi T, Casareto B, Yoshinaga K, Nakano Y *et al*. (2012). Biological and chemical characteristics of the coral gastric cavity. *Coral Reefs* **31:** 147-156.

Ambrose HE, Clewley JP. (2006). Virus discovery by sequence-independent genome amplification. *Rev Med Virol* **16:** 365-383.

Andrews-Pfannkoch C, Fadrosh DW, Thorpe J, Williamson SJ. (2010). Hydroxyapatite-Mediated Separation of Double-Stranded DNA, Single-Stranded DNA, and RNA Genomes from Natural Viral Assemblages. *Appl Environ Microbiol* **76:** 5039-5045.

Angly FE, Felts B, Breitbart M, Salamon P, Edwards RA, Carlson C *et al*. (2006). The marine viromes of four oceanic regions. *PLoS Biol* **4:** e368.

Angly FE, Willner D, Prieto-Davó A, Edwards RA, Schmieder R, Vega-Thurber R *et al*. (2009). The GAAS metagenomic tool and its estimations of viral and microbial average genome size in four major biomes. *PLoS Comput Biol* **5:** e1000593.

Baker ML, Jiang W, Rixon FJ, Chiu W. (2005). Common ancestry of herpesviruses and tailed DNA bacteriophages. *J Virol* **79:** 14967-14970.

Berkelmans R, De'ath G, Kininmonth S, Skirving WJ. (2004). A comparison of the 1998 and 2002 coral bleaching events on the Great Barrier Reef: spatial correlation, patterns, and predictions. *Coral Reefs*

23: 74-83.
Bexfield N, Kellam P. (2011). Metagenomics and the molecular identification of novel viruses. *The Veterinary Journal* **190:** 191-198.
Bigot Y, Renault S, Nicolas J, Moundras C, Demattei M-V, Samain S *et al*. (2009). Symbiotic virus at the evolutionary intersection of three types of large DNA viruses; Iridoviruses, Ascoviruses, and Ichnoviruses. *PLoS ONE* **4:** e6397.
Bourne DG, Munn CB. (2005). Diversity of bacteria associated with the coral *Pocillopora damicornis* from the Great Barrier Reef. *Environ Microbiol* **7:** 1162-1174.
Breitbart M, Salamon P, Andresen B, Mahaffy JM, Segall AM, Mead D *et al*. (2002). Genomic analysis of uncultured marine viral communities. *Proc Natl Acad Sci USA* **99:** 14250-14255.
Breitbart M, Thompson LR, Suttle CA, Sullivan MB. (2007). Exploring the vast diversity of marine viruses. *Oceanography* **20:** 135-139.
Canchaya C, Fournous G, Chibani-Chennoufi S, Dillmann M-L, Brussow H. (2003). Phage as agents of lateral gene transfer. *Curr Opin Microbiol* **6:** 417-424.
Carpenter KE, Abrar M, Aeby G, Aronson RB, Banks S, Bruckner A *et al*. (2008). One-Third of Reef-Building Corals Face Elevated Extinction Risk from Climate Change and Local Impacts. *Science* **321:** 560-563.
Correa AMS, Welsh RM, Vega Thurber RL. (2013). Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. *ISME J* **7:** 13-27.
Culley AI, Suttle CA. (2010). Charcterization of the diversity of marine RNA viruses. In: Wilhelm SW, Weinbauer MG, Suttle CA (eds). *Manual of Aquatic Viral Ecology*. ASLO. pp 193-201.
Davy SK, Burchett SG, Dale AL, Davies P, Davy JE, Muncke C *et al*. (2006). Viruses: agents of coral disease? *Dis Aquat Organ* **69:** 101-110.
De'ath G, Fabricius KE, Sweatman H, Puotinen M. (2012). The 27–year decline of coral cover on the Great Barrier Reef and its causes. *Proc Natl Acad Sci USA* **109:** 17995-17999.
Dinsdale EA, Pantos O, Smriga S, Edwards RA, Angly F, Wegley L *et al*. (2008). Microbial ecology of four coral atolls in the Northern Line Islands. *PLoS ONE* **3:** e1584.
Djikeng A, Halpin R, Kuzmickas R, DePasse J, Feldblyum J, Sengamalay N *et al*. (2008). Viral genome sequencing by random priming methods. *BMC Genomics* **9:** 5.
Ducklow HW, Mitchell R. (1979). Bacterial populations and adaptations in the mucus layers on living corals. *Limnol Oceanogr* **24:** 715.
Duhaime MB, Deng L, Poulos BT, Sullivan MB. (2012). Towards quantitative metagenomics of wild viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization of the linker amplification method. *Environ Microbiol* **14:** 2526-2537.
Edwards RA, Rohwer F. (2005). Viral metagenomics. *Nat Rev Micro* **3:** 504-510.
Fields BN, Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA *et al* (2007). *Fields virology*, 5 edn. Lippincott Williams and Wilkins: Philadelphia, PA.
Garcia-Heredia I, Martin-Cuadrado A-B, Mojica FJM, Santos F, Mira A, Antón J *et al*. (2012). Reconstructing viral genomes from the environment using fosmid clones: the case of Haloviruses. *PLoS ONE* **7:** e33802.
Hanson L, Dishon A, Kotler M. (2011). Herpesviruses that infect fish. *Viruses* **3:** 2160-2191.
Henn MR, Sullivan MB, Stange-Thomann N, Osburne MS, Berlin AM, Kelly L *et al*. (2010). Analysis of high-throughput sequencing and annotation strategies for phage genomes. *PLoS ONE* **5:** e9083.
Hewson I, Brown J, Burge C, Couch C, LaBarre B, Mouchka M *et al*. (2012). Description of viral assemblages associated with the *Gorgonia ventalina* holobiont. *Coral Reefs* **31:** 487-491.
Hick P, Tweedie A, Whittington R. (2010). Preparation of fish tissues for optimal detection of betanodavirus. *Aquaculture* **310:** 20-26.
Hoegh-Guldberg O. (1999). Climate change, coral bleaching, and the future of the world's coral reefs. *Mar Freshw Res* **50:** 839-866.
Homa FL, Brown JC. (1997). Capsid assembly and DNA packaging in herpes simplex virus. *Rev Med Virol* **7:** 107-122.
ICTV (2012). *Virus taxonomy: classification and nomenclature of viruses: Ninth Report of the International Committee on Taxonomy of Viruses*, 9 edn. Elsevier Academic Press: San Diego.
Jacobsen A, Bratbak G, Heldal M. (1996). Isolation and characterization of a virus infecting *Phaeocystis pouchetii* (Prymnesiophyceae). *J Phycol* **32:** 923-927.
Jiang SC, Kellogg CA, Paul JH. (1998). Characterization of marine temperate phage-host systems isolated from Mamala Bay, Oahu, Hawaii. *Appl Environ Microbiol* **64:** 535-542.
Kariin S, Burge C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11:** 283-290.
Kim K-H, Bae J-W. (2011). Amplification methods bias metagenomic libraries of uncultured single-stranded and double-stranded DNA viruses. *Appl Environ Microbiol* **77:** 7663-7668.
Koonin EV, Dolja VV. (2012). Expanding networks of RNA virus evolution. *BMC Biol* **10:** 54.
Kooperman N, Ben-Dov E, Kramarsky-Winter E, Barak Z, Kushmaro A. (2007). Coral mucus-associated bacterial communities from natural and aquarium environments. *FEMS Microbiol Lett* **276:** 106-113.
Kristensen DM, Mushegian AR, Dolja VV, Koonin EV. (2010). New dimensions of the virus world discovered through metagenomics. *Trends Microbiol* **18:** 11-19.
Lang AS, Zhaxybayeva O, Beatty JT. (2012). Gene transfer agents: phage-like elements of genetic exchange. *Nat Rev Micro* **10:** 472-482.
Lawrence JE, Steward GF. (2010). Purification of viruses by centrifugation. In: Wilhelm SW, Weinbauer MG, Suttle CA (eds). *Manual of Aquatic Viral Ecology*. ALSO. pp 166-181.
Littman R, Willis BL, Bourne DG. (2011). Metagenomic analysis of the coral holobiont during a natural bleaching event on the Great Barrier Reef. *Environ Microbiol Rep* **3:** 651-660.
Littman RA, Willis BL, Pfeffer C, Bourne DG. (2009). Diversities of coral-associated bacteria differ with location, but not species, for three acroporid corals on the Great Barrier Reef. *FEMS Microbiol Ecol* **68:** 152-163.
Liu HQ, Fu YP, Li B, Yu X, Xie JT, Cheng JS *et al*. (2011). Widespread horizontal gene transfer from circular single-stranded DNA viruses to eukaryotic genomes. *BMC Evol Biol* **11**.
Lohr J, Munn CB, Wilson WH. (2007). Characterization of a latent virus-like infection of symbiotic zooxanthellae. *Appl Environ Microbiol* **73:** 2976-2981.
Maness HTD, Nollens HH, Jensen ED, Goldstein T, LaMere S, Childress A *et al*. (2011). Phylogenetic analysis of marine mammal herpesviruses. *Vet Microbiol* **149:** 23-29.
Marhaver KL, Edwards RA, Rohwer F. (2008). Viral communities associated with healthy and bleaching corals. *Environ Microbiol* **10:** 2277-2286.
Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M *et al*. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9:** 386.
Monier A, Pagarete An, de Vargas C, Allen MJ, Read B, Claverie J-M *et al*. (2009). Horizontal gene transfer of an entire metabolic pathway between a eukaryotic alga and its DNA virus. *Genome Res* **19:** 1441-1449.
Munn CB. (2006). Viruses as pathogens of marine organisms - from bacteria to whales. *J Mar Biol Assoc U K* **86:** 453-467.
Patten N, Harrison P, Mitchell J. (2008). Prevalence of virus-like particles within a staghorn scleractinian coral (*Acropora muricata*) from the Great Barrier Reef. *Coral Reefs* **27:** 569-580.
Raes J, Bork P. (2008). Molecular eco-systems biology: towards an understanding of community function. *Nat Rev Micro* **6:** 693-699.
Raina J-B, Tapiolas D, Willis BL, Bourne DG. (2009). Coral-associated bacteria and their role in the biogeochemical cycling of sulfur. *Appl Environ Microbiol* **75:** 3492-3501.
Ritchie KB, Smith GW. (2004). Microbial communities of coral surface mucopolysaccharide layers. In: Rosenberg E, Loya Y (eds). *Coral health and disease*. Springer-Verlag: Berlin, Germany. pp 259-263.
Rohwer F, Breitbart M, Jara J, Azam F, Knowlton N. (2001). Diversity of bacteria associated with the Caribbean coral *Montastraea franksi*. *Coral Reefs* **20:** 85-91.
Rohwer F, Seguritan V, Azam F, Knowlton N. (2002). Diversity and distribution of coral-associated bacteria. *Mar Ecol Prog Ser* **243:** 10.
Roossinck MJ. (2011). The good viruses: viral mutualistic symbioses. *Nat Rev Micro* **9:** 99-108.
Rosario K, Breitbart M. (2011). Exploring the viral world through metagenomics. *Curr Opin Virology* **1:** 1-9.
Rosario K, Duffy S, Breitbart M. (2012). A field guide to eukaryotic circular single-stranded DNA viruses: insights gained from metagenomics. *Arch Virol***:** 1-21.
Rosenberg E, Koren O, Reshef L, Efrony R, Zilber-Rosenberg I. (2007). The role of microorganisms in coral health, disease and evolution. *Nat Rev Microbiol* **5:** 355-362.
Roux S, Faubladier M, Mahul A, Paulhe N, Bernard A, Debroas D *et al*. (2011). Metavir: a web server dedicated to virome analysis. *Bioinformatics* **27:** 3074-3075.
Roux S, Enault F, Robin A, Ravet V, Personnic S, Theil S *et al*. (2012). Assessing the Diversity and Specificity of Two Freshwater Viral Communities through Metagenomics. *PLoS ONE* **7:** e33641.
Sambrook J, Russell DW (2001). *Molecular cloning: A laboratory manual*, 3rd edn. Cold Springs Harbor Laboratory Press: Harbour, New York.

Santos F, Yarza P, Parro V, Briones C, Antón J. (2010). The metavirome of a hypersaline environment. *Environ Microbiol* **12:** 2965-2976.

Stasiak K, Renault S, Demattei M-V, Bigot Y, Federici BA. (2003). Evidence for the evolution of ascoviruses from iridoviruses. *J Gen Virol* **84:** 2999-3009.

Steward GF, Preston CM. (2011). Analysis of a viral metagenomic library from 200m depth in Monterey Bay, California constructed by direct shotgun cloning. *Virol J* **8:** 14.

Steward GF, Culley AI, Mueller JA, Wood-Charlson EM, Belcaid M, Poisson G. (2012). Are we missing half of the viruses in the ocean? *ISME J*.

Steward GF, Culley AI, Wood-Charlson EM. (2013). Marine Viruses. In: Editor-in-Chief: Simon AL (ed). *Encyclopedia of Biodiversity (Second Edition)*. Academic Press: Waltham. pp 127-144.

Sweatman H, Delean S, Syms C. (2011). Assessing loss of coral cover on Australia's Great Barrier Reef over two decades, with implications for longer-term trends. *Coral Reefs* **30:** 521-531.

Sweet M, Croquer A, Bythell J. (2011). Bacterial assemblages differ between compartments within the coral holobiont. *Coral Reefs* **30:** 39-52.

Thurber RLV, Correa AMS. (2011). Viruses of reef-building scleractinian corals. *J Exp Mar Biol Ecol* **408:** 102-113.

van Oppen M, Leong J-A, Gates RD. (2009). Coral-virus interactions: A double-edged sword? *Symbiosis* **47:** 1-8.

Vega Thurber RL, Barott KL, Hall D, Liu H, Rodriguez-Mueller B, Desnues C *et al*. (2008). Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral *Porites compressa*. *Proc Natl Acad Sci USA* **105:** 18413-18418.

Vega Thurber RV, Haynes M, Breitbart M, Wegley L, Rohwer F. (2009). Laboratory procedures to generate viral metagenomes. *Nat Protocols* **4:** 470-483.

Wegley L, Edwards R, Rodriguez-Brito B, Liu H, Rohwer F. (2007). Metagenomic analysis of the microbial community associated with the coral *Porites astreoides*. *Environ Microbiol* **9:** 2707-2719.

Wellington GM, Glynn PW, Strong AE, Navarrete SA, Wieters E, Hubbard D. (2001). Crisis on coral reefs linked to climate change. *Eos Trans AGU* **82:** 1-5.

Weston AJ, Dunlap WC, Shick JM, Klueter A, Iglic K, Vukelic A *et al*. (2012). A Profile of an Endosymbiont-enriched Fraction of the Coral Stylophora pistillata Reveals Proteins Relevant to Microbial-Host Interactions. *Mol Cell Proteomics* **11**.

Weynberg KD, Allen MJ, Ashelford K, Scanlan DJ, Wilson WH. (2009). From small hosts come big viruses: the complete genome of a second *Ostreococcus tauri* virus, OtV-1. *Environ Microbiol* **11:** 2821-2839.

Weynberg KD, Allen MJ, Gilg IC, Scanlan DJ, Wilson WH. (2011). Genome sequence of *Ostreococcus tauri* virus OtV-2 throws light on the role of picoeukaryote niche separation in the ocean. *J Virol* **85:** 4520-4529.

Williamson SJ, Rusch DB, Yooseph S, Halpern AL, Heidelberg KB, Glass JI *et al*. (2008). The Sorcerer II Global Ocean Sampling Expedition: Metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE* **3:** e1456.

Willner D, Thurber RV, Rohwer F. (2009). Metagenomic signatures of 86 microbial and viral metagenomes. *Environ Microbiol* **11:** 1752-1766.

Wilson WH, Francis I, Ryan K, Davy SK. (2001). Temperature induction of viruses in symbiotic dinoflagellates. *Aquat Microb Ecol* **25:** 99-102.

Wilson WH, Dale AL, Davy JE, Davy SK. (2005). An enemy within? Observations of virus-like particles in reef corals. *Coral Reefs* **24:** 145-148.

Winter A, Appeldoorn RS, Bruckner A, Williams Jr. EH, Goenaga C. (1998). Sea surface temperatures and coral reef bleaching off La Parguera, Puerto Rico (northeastern Caribbean Sea). *Coral Reefs* **17:** 377-382.

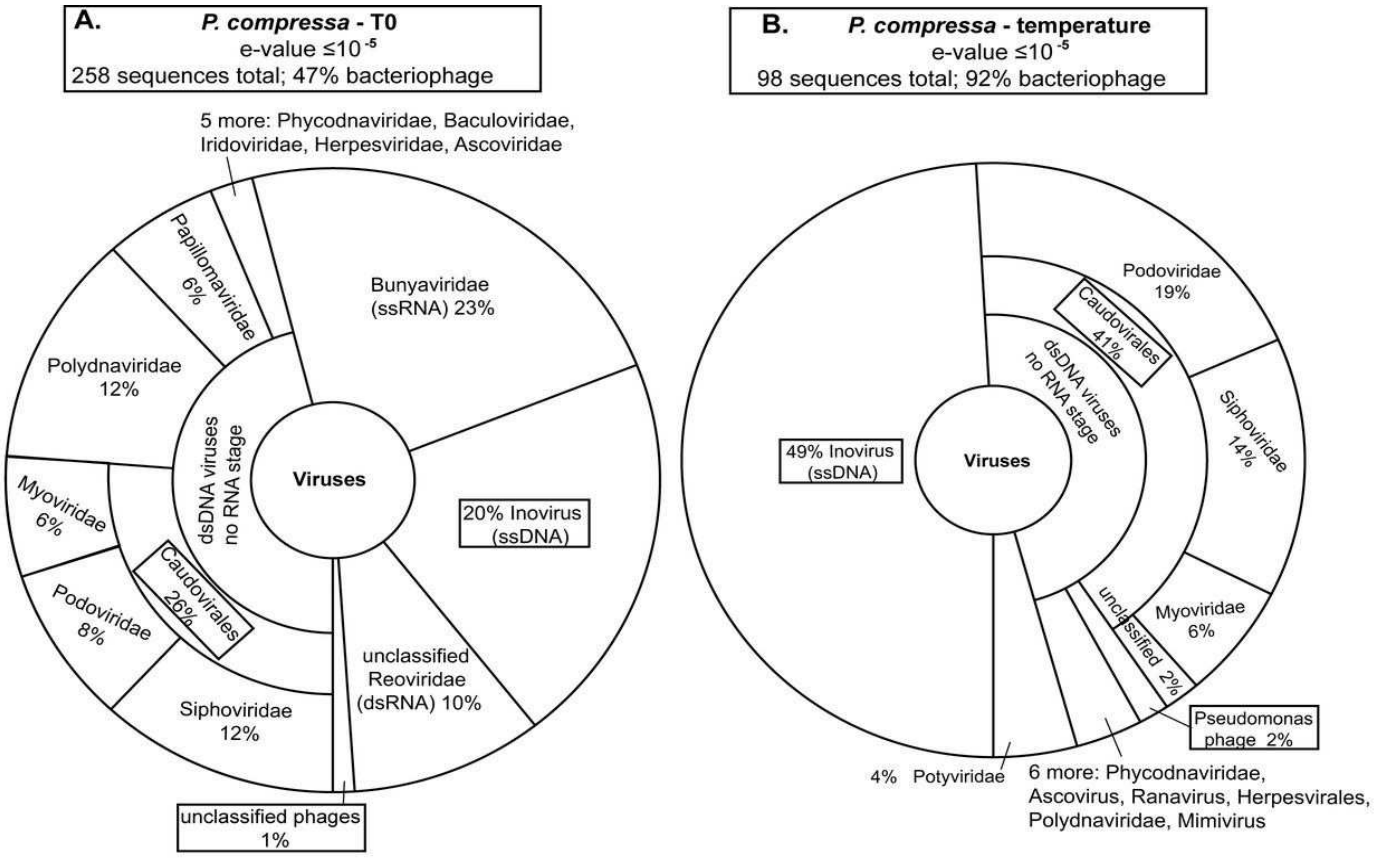
Wommack KE, Bhavsar J, Ravel J. (2008). Metagenomics: Read length matters. *Appl Environ Microbiol* **74:** 1453-1463.

Wommack KE, Bhavsar J, Polson SW, Chen J, Dumas M, Srinivasiah S *et al* (2012). *VIROME: a standard operating procedure for analysis of viral metagenome sequences*, vol. 6.

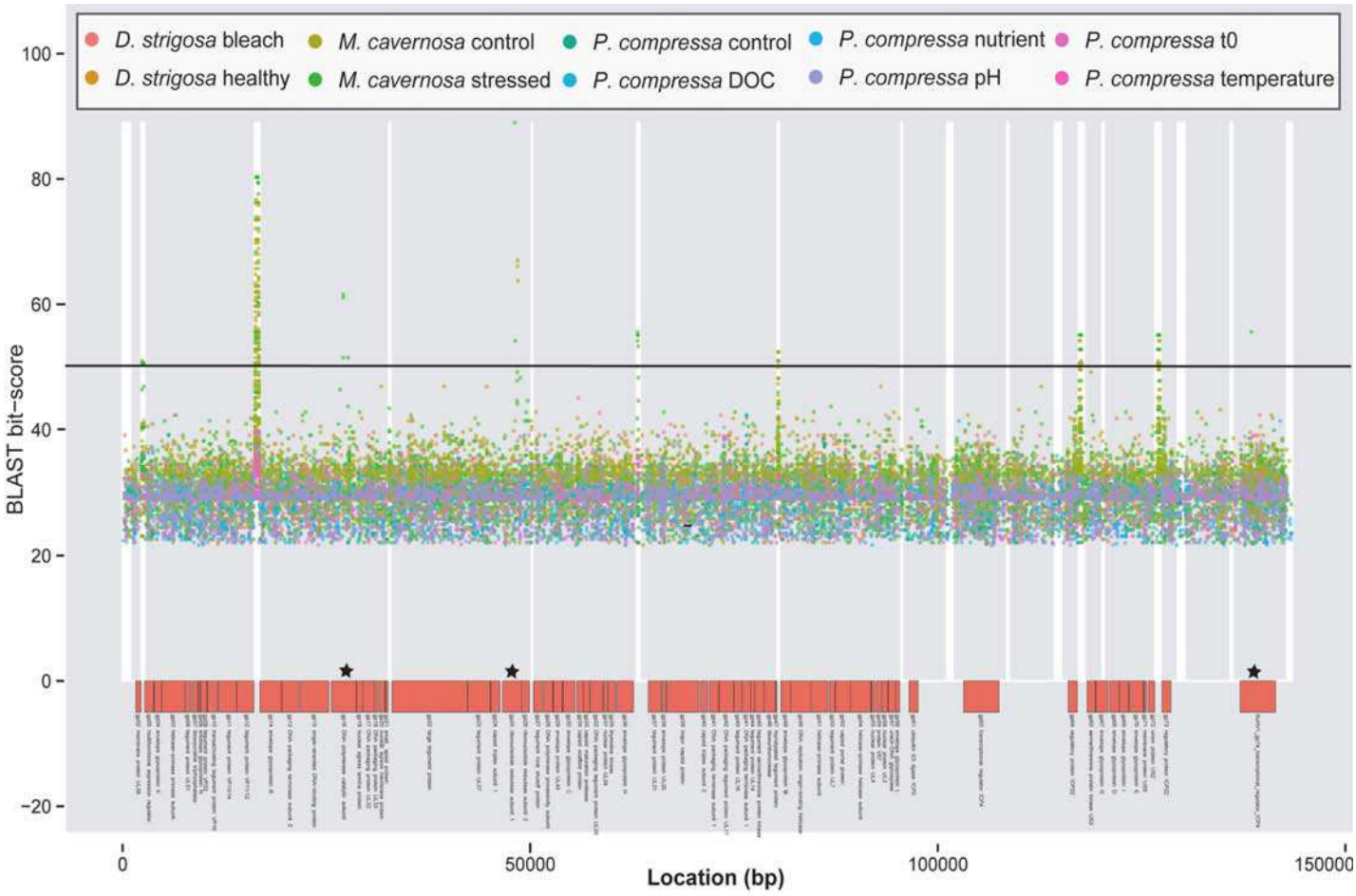
Xiao C, Chipman PR, Battisti AJ, Bowman VD, Renesto P, Raoult D *et al*. (2005). Cryo-electron microscopy of the giant mimivirus. *J Mol Biol* **353:** 493-496.

Yildiz FH, Visick KL. (2009). Vibrio biofilms: so much the same yet so different. *Trends Microbiol* **17:** 109-118.

Yokouchi H, Fukuoka Y, Mukoyama D, Calugay R, Takeyama H, Matsunaga T. (2006). Whole-metagenome amplification of a microbial community associated with scleractinian coral by multiple displacement amplification using ϕ29 polymerase. *Environ Microbiol* **8:** 1155-1163.



Supplemental Figure 1. Viral composition (e-value ≤10-5) for the Porites compressa A) time zero (T0) and B) temperature treatments. Boxed labels represent viruses that infect bacteria.



Supplemental Figure 2. Recruitment plot of the virome data sets from Porites compressa and Montastraea cavernosa against the Suid herpesvirus 1 genome (SuHV1, NCBI genome NC_006151). Annotated regions are represented by the red blocks, and all repeat regions within non-coding regions are highlighted in white. BLAST bit-scores (size independent thresholds for database comparison) greater than 50 (area above black line) is generally considered a significant match. Stars mark annotated proteins with significant bit-scores.

Supplementary Table 1. List of metagenomes and transcriptomes used in this study.

Species	Research Lab	Sequence type	Holobiont fraction	Holobiont development stage	Online database for original sequence files	Chloroform (y/n); Amplification method	Sequencing platform	Citation
Acropora hyacinthus	Matz, M.	Transcriptome – v.1 Jul 2010	Coral	Larvae	http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html		454	
		Transcriptome – v.1 Jan 2012	Coral	Larvae	http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html		454	
Acropora millepora	Matz, M.	Transcriptome – v.3 Jul 2010	Coral	Larvae	http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html		454	Meyer, E. *et al.* (2009)
	Matz, M.	Transcriptome – v. Jan2012	Coral	Adult, larvae	http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html		454	
	Bourne, D.	Metagenome – Pre-bleach	Prokaryote	Adult	http://metagenomics.anl.gov/linkin.cgi?metagenome=4445756.3	No; MDA[1]	454	Littman, R. *et al.* (2011)
	Bourne, D.	Metagenome – Post-bleach	Prokaryote	Adult	http://metagenomics.anl.gov/linkin.cgi?metagenome=4445755.3		454	
Acropora palmata	Baums, I.	Transcriptome – 2011	Coral	Adult, larvae	http://main.g2.bx.psu.edu/u/nickpolato/h/apalmataassembly		454	Polato, N.R. *et al.* (2011)
Acropora tenuis	Bay, L.	Transcriptome – v. Jan 2011	Coral	Larvae, recruits	http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html		454	
Diploria strigosa	Rohwer, F.	Metagenome – Healthy	Virus	Adult	http://www.ncbi.nlm.nih.gov/nuccore/ABVU00000000.1	Yes; LASL[2]	capillary	Marhaver, K.L. *et al.* (2008)
	Rohwer, F.	Metagenome – Bleached	Virus	Adult	http://www.ncbi.nlm.nih.gov/nuccore/ABVT00000000.1			
Montastraea annularis	Medina, M.	Transcriptome	Coral	Adult, larvae	http://sequoia.ucmerced.edu/SymBioSys/index.php		capillary	
Montastraea cavernosa	Vega Thurber, R.	Metagenome – Control	Virus	Adult	http://metagenomics.anl.gov/linkin.cgi?metagenome=4455158.3	No; WTA[3]	454	Correa, A.M.S. *et al.* (2012)
		Metagenome – Temperature	Virus	Adult	http://metagenomics.anl.gov/linkin.cgi?metagenome=4455159.3			
Montastraea faveolata	Medina, M.	Transcriptome	Coral	Adult, larvae	http://sequoia.ucmerced.edu/SymBioSys/index.php		capillary	Schwarz, J.A. *et al.* (2008)
Porites astreoides	Matz, M.	Transcriptome – v.1 Jul 2010	Coral	Adult	http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html		454	
	Matz, M.	Transcriptome – v. Jan 2012	Coral	Adult, larvae	http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html		454	
	Rohwer, F.	Metagenome	Prokaryote	Adult	http://metagenomics.anl.gov/linkin.cgi?metagenome=4440319.3	No; MDA	454	Wegley, L. *et al.* (2007)
Porites compressa	Vega Thurber, R.	Metagenome – 6 treatments	Virus	Adult	http://metavir-meb.univ-bpclermont.fr/ "P. compressa - VegaThurber et al., 2008"	Yes; MDA	454	Vega Thurber, R.L. *et al.* (2008)
Pocillipora damicornis	van Oppen, M.	Metagenome – DNA Metagenome – RNA	Virus	Adult	http://metagenomics.anl.gov/linkin.cgi?metagenome=4492317.3 http://metagenomics.anl.gov/linkin.cgi?metagenome=4492316.3	Yes; MDA	454	
Coral reef seawater	Rohwer, F.	Metagenome – Seawater from 4 coral atolls in Line Islands	Virus	n/a	http://metavir-meb.univ-bpclermont.fr/ "Coral Atoll - Dinsdale et al., 2008"	Yes; MDA	454	Dinsdale, E.A. *et al.* (2008)

[1] MDA: multi-displacement amplification
[2] LASL: Linker-amplified shotgun library
[3] WTA: whole transcriptome amplification

References:

Correa AM, Welsh RM and Vega Thurber RL (2012) Unique nucleocytoplasmic dsDNA and +ssRNA viruses are associated with the dinoflagellate endosymbionts of corals. ISME J.
Dinsdale EA, et al. (2008) Microbial ecology of four coral atolls in the Northern Line Islands. PLoS ONE 3(2): e1584.
Littman R, Willis BL and Bourne DG (2011) Metagenomic analysis of the coral holobiont during a natural bleaching event on the Great Barrier Reef. Environ Microbiol Rep 3(6): 651-660.
Marhaver KL, Edwards RA and Rohwer F (2008) Viral communities associated with healthy and bleaching corals. Environ Microbiol 10(9): 2277-2286.
Meyer E, et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GS-Flx. BMC Genomics 10(1): 219.
Polato NR, Vera JC and Baums IB (2011) Gene discovery in the threatened Elkhorn Coral: 454 sequencing of the Acropora palmata transcriptome. PLoS ONE 6(12): e28634.
Schwarz JA, et al. (2008) Coral life history and symbiosis: functional genomic resources for two reef building Caribbean corals, Acropora palmata and Montastraea faveolata. BMC Genomics 9: 97.
Vega Thurber RL, et al. (2008) Metagenomic analysis indicates that stressors induce production of herpes-like viruses in the coral Porites compressa. Proc Natl Acad Sci USA 105(47): 18413-18418.
Wegley L, Edwards R, Rodriguez-Brito B, Liu H and Rohwer F (2007) Metagenomic analysis of the microbial community associated with the coral Porites astreoides. Environ Microbiol 9(11): 2707-2719.

Supplementary Table 2. Community composition of coral-associated viruses, as determined by genome chemistry, in the complete data sets.

		Percent of viruses out of whole (e-value 10^{-5})								
Species	**Sequence type**	**RT**	**dsDNA, no RNA**	**ssDNA**	**dsRNA**	**ssRNA+, No DNA**	**ssRNA-**	**Archaeal**	**Unclassified**	**Satellite**
Acropora hyacinthus	Transcriptome – 2010	75.50	20.90	0.40	1.90	0.90	0.20	0.10	0.10	
	Transcriptome – 2012	74.80	21.00		2.80	1.00		0.20	0.20	
Acropora millepora	Transcriptome – 2010	72.00	22.50		4.50	0.50	0.20	0.10	0.20	
	Transcriptome – 2012	66.00	27.60		4.90	0.70	0.40		0.40	
	Metagenome – Pre-bleach	62.70	9.30	27.70	0.20				0.10	
	Metagenome – Post-bleach	37.30	3.70	57.20	0.03	0.20				1.60
Acropora palmata	Transcriptome – 2011	58.30	31.60	0.70	8.00	0.80	0.10		0.50	
Acropora tenuis	Transcriptome – 2012	74.10	19.20	0.60	4.80	0.50	0.40		0.40	
Diploria strigosa	Virome – Healthy	2.40	95.40						2.20	
	Virome – Bleached		98.80						1.20	
Montastrea annularis	Transcriptome	59.80	22.30		17.60					0.30
Montastraea cavernosa	Virome – Control	4.40	11.40	83.30			0.10		0.80	
	Virome – Temperature	4.30	3.60	89.50	0.10		0.60		1.30	0.60
Montastrea faveolata	Transcriptome	59.80	35.40		1.10	3.10	0.40		0.20	
Porites astreoides	Transcriptome – 2010	52.70	35.40		4.00	7.40			0.50	
	Transcriptome – 2012	55.50	28.90		5.10	9.70	0.20		0.60	
	Metagenome – Prokaryote		3.90	95.70	0.40				0.04	
Porites compressa	Virome – Control		29.20	69.00	0.20	1.00			0.60	
	Virome – DOC		22.60	76.40					1.00	
	Virome – Nutrient		65.70	27.50					6.80	
	Virome – pH	1.60	78.70	9.80			0.90		9.00	
	Virome – Time 0		46.10	20.40	9.50		23.00		1.00	
	Virome – Temperature		46.60	49.10		4.30				
Pocillipora damicornis	Virome – DNA	0.10	82.00	14.80		0.10			2.50	0.50
	Virome – RNA		7.50	88.20	0.02	3.30			0.80	0.20
Coral atolls	Virome – Seawater									
Kingman Island			98.40						1.60	
Palmyra Island			95.90	2.50		0.50	0.10		1.00	
Tabueran Island			96.30	3.20	0.07				0.50	
Kiritimati Island			91.50	3.70		0.80			4.00	

Annexe A.4 : Matériel supplémentaire

Supplementary material : Uncontaminated viromes reveal the abundance and diversity of metabolism genes in environmental viruses (Article III)

Table S1 : **List of viromes and microbiomes used in this study.** The web-servers hosting the datasets are: NCBI (www.ncbi.nlm.nih.gov), MG-Rast (http://metagenomics.anl.gov), and Metavir (http://metavir-meb.univ-bpclermont.fr). Each virome and microbiome is identified by an Id. throughout the paper. When available, the methodology used to purify viral particle is indicated (CsCl : Cesium Chloride, PEG : Polyethylene Glycol, LASL : linker amplified shotgun library and MDA : phi29-mediated multiple displacement amplification).

Virome id	Virome name	Available on	Methodology used in sample preparation	Sample origin	Sample type	Number of sequences	Mean size of sequences
12	Medium viruses (MP1128)	MG-Rast – 4440427.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	39,439	100.43
13	Medium viruses (MP1116)	MG-Rast – 4440428.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	58,319	98.11
14	High saltern viral (HP1116)	MG-Rast – 4440421.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	151,180	99.76
15	Low saltern (Pond 11) viruses	MG-Rast – 4440436.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	268,049	104.47
16	Low saltern (Pond 11) viruses (LP1110)	MG-Rast – 4440432.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	109,836	104.35
17	Medium saltern viruses (Pond MP1110)	MG-Rast – 4440431.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	39,348	101.40
18	Medium saltern viruses (MP1122)	MG-Rast – 4440417.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	55,142	100.79
19	High saltern viruses (HP1207)	MG-Rast – 4440145.4	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	46,628	102.15
20	High saltern viral (HP1128)	MG-Rast – 4440144.4	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	4,536	100.15
21	Low saltern viruses (LP1128)	MG-Rast – 4440420.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	62,363	103.87
22	Salton Sea Phage 1	MG-Rast – 4440327.3		Salton Sea – California	Hypersaline	55,467	103.60
23	Salton Sea Phage 3	MG-Rast – 4440328.3		Salton Sea – California	Hypersaline	29,814	99.20
32	Marine Phages GOM	MG-Rast – 4440304.3	CsCL – MDA	Gulf of Mexico	Seawater	262,501	101.59
33	Marine Phages BBC	MG-Rast – 4440305.3	CsCL – MDA	British Columbia	Seawater	414,964	102.14
34	Marine Phages Arctic	MG-Rast – 4440306.3	CsCL – MDA	Arctic sea	Seawater	686,209	99.15
35	Marine Phage SAR	MG-Rast – 4440322.3	CsCL – MDA	Sargasso sea	Seawater	397,939	104.31
36	Line Islands Kingman Reef B2 phage	MG-Rast – 4440036.3	CsCL – MDA	Kingmann - Line Islands	Seawater	93,744	108.34
37	Line Islands Christmas Reef B3 phage	MG-Rast – 4440038.3	CsCL – MDA	Christmas - Line Islands	Seawater	279,882	110.56
38	Line Islands Palmyra F8 Phage	MG-Rast – 4440040.3	CsCL – MDA	Palmyra - Line Islands	Seawater	318,178	104.78
39	Line Islands Tabuaeran B1 Phage	MG-Rast – 4440280.3	CsCL – MDA	Tabuaeran - Line Islands	Seawater	378,475	104.13
40	Tampa Bay phage from induction experiment	MG-Rast – 4440102.3	PEG – CsCL- MDA	Tampa Bay – Florida	Seawater	279,129	103.95
41	Skan Bay Phage 1	MG-Rast – 4440330.3		Skan Bay – Alaska	Seawater	30,831	104.56
46	Tpond phage 3	MG-Rast – 4440424.3	PEG – CsCL- MDA	Tilapia Pond 3 – California	Freshwater	56,549	101.06
47	Healthy Tilapia pond phages	MG-Rast – 4440412.3	PEG – CsCL- MDA	Healthy fish Pond – California	Freshwater	60,135	101.07
48	Healthy Prebead tank phages	MG-Rast – 4440414.3	PEG – CsCL- MDA	Prebead Pond – California	Freshwater	67,785	103.21
49	Tilapia pond	MG-Rast – 4440439.3	PEG – CsCL- MDA	Tilapia Pond – California	Freshwater	264,844	102.25
57	Porites compressa time zero viruses	MG-Rast – 4440376.3	CsCL – MDA	Porites Compressa	Eukaryote	39,113	101.32
58	Porites compressa control treated viruses	MG-Rast – 4440374.3	CsCL – MDA	Porites Compressa	Eukaryote	39,191	103.70
59	Porites compressa DOC treated viruses	MG-Rast – 4440370.3	CsCL – MDA	Porites Compressa	Eukaryote	35,409	102.18
60	Porites compressa pH treated viruses	MG-Rast – 4440371.3	CsCL – MDA	Porites Compressa	Eukaryote	49,949	104.73
61	Porites compressa nutrient treated viruses	MG-Rast – 4440377.3	CsCL – MDA	Porites Compressa	Eukaryote	34,139	107.18
62	Porites compressa temperature treated viruses	MG-Rast – 4440375.3	CsCL – MDA	Porites Compressa	Eukaryote	38,482	113.38
66	Pozas Azules Stromatolites phages	MG-Rast – 4440320.3	CsCL – MDA	Paztac Azules	Microbialites	301,264	104.64
67	Rios Mesquites Stromatolites phages	MG-Rast – 4440321.3	CsCL – MDA	Rio Mesquites	Microbialites	324,500	104.22
68	Highborne Cay Stromatolite phage	MG-Rast – 4440323.3	CsCL – MDA	Bahamas	Microbialites	148,334	100.52
73	Healthy slime viruses	MG-Rast – 4440065.3		Healthy fish gut	Eukaryote	61,022	98.45
74	Morbid slime viruses	MG-Rast – 4440064.3		Morbid fish gut	Eukaryote	59,599	98.32
83	CF Lung Sputum Viruses	MG-Rast – 4440441.3	CsCL – MDA	Human Lung (USA)	Eukaryote	92,223	80.71
84	Health Lung Sputum Viruses	MG-Rast – 4440442.4	CsCL – MDA	Human Lung (USA)	Eukaryote	39,489	84.80
85	Mosquito DNA 1	MG-Rast – 4440052.3	CsCL – MDA	Mosquito (USA)	Eukaryote	336,760	102.61
86	Mosquito DNA 2	MG-Rast – 4440053.3	CsCL – MDA	Mosquito (USA)	Eukaryote	638,689	100.32
87	Mosquito DNA 3	MG-Rast – 4440054.3	CsCL – MDA	Mosquito (USA)	Eukaryote	601,040	104.16
AISp	Lake Limnopolar Spring	Metavir – Project Lake Limnopolar	Sucrose cushion – MDA	Lake Limnopolar	Freshwater	41,322	239.65
AISu	Lake Limnopolar Summer	Metavir – Project Lake Limnopolar	Sucrose cushion – MDA	Lake Limnopolar	Freshwater	38,475	221.27
MET6	Lake Pavin	Metavir – Project French Lakes	PEG – MDA	Lake Pavin – France	Freshwater	649,290	412.31
MET7	Lake Bourget	Metavir – Project French Lakes	PEG – MDA	Lake Bourget – France	Freshwater	593,084	433.41
SRR043421	Human Gut L1 8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – X1	Eukaryote	30,873	372.90
SRR043422	Human Gut L2 1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L1	Eukaryote	132,569	379.12
SRR043423	Human Gut L2 7	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	61,104	370.82
SRR043424	Human Gut L2 8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	148,781	370.31
SRR043425	Human Gut H1 7	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	16,955	375.02
SRR043426	Human Gut H1 8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	13,048	378.90
SRR043427	Human Gut H2 8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	16,747	365.41
SRR043428	Human Gut L1 1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H2	Eukaryote	16,137	366.71
SRR043430	Human Gut H1 2	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L1	Eukaryote	107,259	405.06
SRR043431	Human Gut H1 1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	107,993	409.48
SRR043432	Human Gut H2 1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	25,648	377.67
SRR043433	Human Gut L3 1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H2	Eukaryote	23,614	372.91
SRR043434	Human Gut L3 2	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	33,489	369.66
SRR043435	Human Gut L3 7	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	76,090	384.09
SRR043436	Human Gut L3 8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	15,166	382.16
SRR043437	Human Gut F-A	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	59,155	382.72
SRR089800	Human Gut F-B	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	113,054	431.05
SRR089802	Human Gut F-C	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	109,569	435.21
SRR089803	Human Gut F-D	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	68,391	437.07
SRR089804	Human Gut F-E	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	115,121	433.08
SRR089805	Human Gut Unknown	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	98,511	417.05

Next page :

Table S2 : **rDNA ratios, Microbial hit ratio (MHR) and prophage hit ratio (PHR) for each metagenome**, with the different data used to calculate these ratios (total number of sequences, number of 16S and 23S rDNA detected, number of sequences with a microbial hits, number of microbial hits within a prophage-like region). rDNA ratios are colored as in Figure 1, *i.e.* green for 0, orange for 0 to 2 ‰, and red for more than 2‰. MHR are highlighted in blue when lower than 5%. PHR are highlighted in red when lower than 10% and in green when greater than 25%.

Virome Id	Name	Biome	Total number of sequences	Number of 16S rDNA detected	Number of 23S rDNA detected	rDNA ‰	Total number of hits with prokaryote genomes	Number of hits in prophage-like region	Number of hits in non-prophage region	Microbial Hit Ratio	Prophage Hit Ratio
12	Medium viruses (MP1128)	Hypersaline	39 439	1	2	0.76	360	198	162	0.91%	55.00%
13	Medium viruses (MP1116)	Hypersaline	58 319	11	32	7.37	1 621	345	1 276	2.78%	21.28%
14	High saltern viral (HP1116)	Hypersaline	151 180	3	3	0.40	3 520	1 623	1 897	2.33%	46.11%
15	Low saltern (Pond 11) viruses	Hypersaline	268 049	9	21	1.12	6 847	3 374	3 473	2.55%	49.28%
16	Low saltern (Pond 11) viruses (LP1110)	Hypersaline	109 836	0	1	0.09	2 239	1 195	1 044	2.04%	53.37%
17	Medium saltern viruses (Pond MP1110)	Hypersaline	39 348	0	0	0.00	893	463	430	2.27%	51.85%
18	Medium saltern viruses (MP1122)	Hypersaline	55 142	3	0	0.54	865	513	352	1.57%	59.31%
19	High saltern viruses (HP1207)	Hypersaline	46 628	5	4	1.93	2 606	274	2 332	5.59%	10.51%
20	High saltern viral (HP1128)	Hypersaline	4 536	2	22	52.91	752	98	654	16.58%	13.03%
21	Low saltern viruses (LP1128)	Hypersaline	62 363	63	112	28.06	10 517	997	9 520	16.86%	9.48%
22	Salton Sea Phage 1	Hypersaline	55 467	0	0	0.00	959	546	413	1.73%	56.93%
23	Salton Sea Phage 3	Hypersaline	29 814	0	0	0.00	464	242	222	1.56%	52.16%
32	Marine Phages GOM	Seaw ater	262 501	29	50	3.01	30 214	2 789	27 425	11.51%	9.23%
33	Marine Phages BBC	Seaw ater	414 964	27	61	2.12	21 048	3 762	17 286	5.07%	17.87%
34	Marine Phages Arctic	Seaw ater	686 209	162	231	5.73	202 106	12 898	189 208	29.45%	6.38%
35	Marine Phage SAR	Seaw ater	397 939	7	18	0.63	7 539	2 211	5 328	1.89%	29.33%
36	Line Islands Kingman Reef B2 phage	Seaw ater	93 744	0	8	0.85	5 477	513	4 964	5.84%	9.37%
37	Line Islands Christmas Reef B3 phage	Seaw ater	279 882	52	103	5.54	63 629	2 869	60 760	22.73%	4.51%
38	Line Islands Palmyra F8 Phage	Seaw ater	318 178	0	13	0.41	4 079	1 273	2 806	1.28%	31.21%
39	Line Islands Tabuaeran B1 Phage	Seaw ater	378 475	0	0	0.00	4 533	1 289	3 244	1.20%	28.44%
40	Tampa Bay phage from induction experiment	Seaw ater	279 129	0	0	0.00	5 080	2 476	2 604	1.82%	48.74%
41	Skan Bay Phage 1	Seaw ater	30 831	0	0	0.00	476	209	267	1.54%	43.91%
46	Tpond phage 3	Freshw ater	56 549	0	6	1.06	1 274	810	464	2.25%	63.58%
47	Healthy Tilapia pond phages	Freshw ater	60 135	2	0	0.33	1 352	853	499	2.25%	63.09%
48	Healthy Prebead tank phages	Freshw ater	67 785	2	5	1.03	2 058	1 377	681	3.04%	66.91%
49	Tilapia pond	Freshw ater	264 844	2	9	0.42	9 674	5 999	3 675	3.65%	62.01%
57	Porites compressa time zero viruses	Coral	39 113	2	10	3.07	2412	207	2 205	6.17%	8.58%
58	Porites compressa control treated viruses	Coral	39 191	144	33	45.16	6191	792	5 399	15.80%	12.79%
59	Porites compressa DOC treated viruses	Coral	35 409	18	52	19.77	2126	334	1 792	6.00%	15.71%
60	Porites compressa pH treated viruses	Coral	49 949	2	31	6.61	2505	386	2 119	5.02%	15.41%
61	Porites compressa nutrient treated viruses	Coral	34 139	1	20	6.15	2639	376	2 263	7.73%	14.25%
62	Porites compressa temperature treated viruses	Coral	38 482	16	38	14.03	1 911	196	1 715	4.97%	10.26%
66	Pozas Azules Stromatolites phages	Microbialites	301 264	4	7	0.37	4 361	1 592	2 769	1.45%	36.51%
67	Rios Mesquites Stromatolites phages	Microbialites	324 500	4	8	0.37	11 966	10 158	1 808	3.69%	84.89%
68	Highborne Cay Stromatolite phage	Microbialites	148 334	0	0	0.00	466	184	282	0.31%	39.48%
73	Healthy slime viruses	Fish	61 022	22	89	18.19	12 924	877	12 047	21.18%	6.79%
74	Morbid slime viruses	Fish	59 599	25	82	17.95	20 446	1 194	19 252	34.31%	5.84%
83	CF Lung Sputum Viruses	Human Lung	92 223	6	10	1.73	1 331	129	1 202	1.44%	9.69%
84	Health Lung Sputum Viruses	Human Lung	39 489	0	3	0.76	292	24	268	0.74%	8.22%
85	Mosquito DNA 1	Mosquito	336 760	61	63	3.68	87 878	18 256	69 622	26.10%	20.77%
86	Mosquito DNA 2	Mosquito	638 689	133	114	3.87	257 424	19 480	237 944	40.31%	7.57%
87	Mosquito DNA 3	Mosquito	601 040	162	228	6.49	178 516	20 504	158 012	29.70%	11.49%
AISp	Lake Limnopolar Spring	Freshw ater	41 322	0	2	0.48	640	311	329	0.24%	29.00%
AISu	Lake Limnopolar Summer	Freshw ater	38 475	1	2	0.78	4 457	786	3 671	3.14%	11.52%
MET6	Lake Pavin	Freshw ater	649 290	0	0	0.00	41 460	25 795	15 665	0.26%	62.42%
MET7	Lake Bourget	Freshw ater	593 084	0	0	0.00	96 717	58 217	38 500	0.62%	54.42%
SRR043421	Human Gut L1 8	Human gut	30 873	4	8	3.89	14 767	5 859	8 908	15.77%	23.63%
SRR043422	Human Gut L2 1	Human gut	132 569	13	12	1.89	21 405	14 266	7 139	2.83%	60.41%
SRR043423	Human Gut L2 7	Human gut	61 104	1	2	0.49	6 069	3 203	2 866	1.47%	52.73%
SRR043424	Human Gut L2 8	Human gut	148 781	8	9	1.14	35 356	22 808	12 548	2.02%	58.79%
SRR043425	Human Gut H1 7	Human gut	16 955	1	2	1.77	3 692	2 247	1 445	2.11%	57.98%
SRR043426	Human Gut H1 8	Human gut	13 048	0	1	0.77	5 298	3 524	1 774	3.56%	56.34%
SRR043427	Human Gut H2 8	Human gut	16 747	0	1	0.60	4 763	3 083	1 680	2.97%	58.43%
SRR043428	Human Gut L1 1	Human gut	16 137	0	1	0.62	3 429	2 247	1 182	2.21%	60.78%
SRR043430	Human Gut H1 2	Human gut	107 259	2	12	1.31	39 254	27 821	11 433	3.70%	55.44%
SRR043431	Human Gut H1 1	Human gut	107 993	3	3	0.56	34 802	24 199	10 603	3.02%	65.28%
SRR043432	Human Gut H2 1	Human gut	25 648	0	2	0.78	5 999	3 731	2 268	2.64%	42.18%
SRR043433	Human Gut L3 1	Human gut	23 614	0	1	0.42	2 785	2 027	758	1.58%	70.43%
SRR043434	Human Gut L3 2	Human gut	33 489	1	0	0.30	4 785	3 281	1 504	1.59%	56.87%
SRR043435	Human Gut L3 7	Human gut	76 090	3	8	1.45	14 467	7 963	6 504	3.74%	29.84%
SRR043436	Human Gut L3 8	Human gut	15 166	1	0	0.66	2 722	1 661	1 061	2.29%	42.36%
SRR043437	Human Gut F-A	Human gut	59 155	2	5	1.18	10 975	6 311	4 664	4.41%	34.28%
SRR089800	Human Gut F-B	Human gut	113 054	9	0	0.80	6 308	3 504	2 804	0.25%	59.44%
SRR089802	Human Gut F-C	Human gut	109 569	3	1	0.37	11 199	5 706	5 493	0.88%	63.86%
SRR089803	Human Gut F-D	Human gut	68 391	0	3	0.44	8 597	5 614	2 983	0.55%	61.66%
SRR089804	Human Gut F-E	Human gut	115 121	1	3	0.35	9 733	5 397	4 336	0.91%	34.45%
SRR089805	Human Gut Unknow n	Human gut	98 511	9	5	1.42	5 481	3 384	2 097	0.45%	50.91%

Microbiome Id	Name	Biome	Total number of sequences	Number of 16S rDNA detected	Number of 23S rDNA detected	rDNA ‰	Total number of hits for 10 000 reads	Number of hits in prophage-like region	Number of hits in non-prophage region	Microbial Hit Ratio	Prophage Hit Ratio
1	Soudan Red Stuff	Subterranean	334 386	236	642	26.26	1 230	103	1 127	12.30%	8.37%
2	Soudan Black Stuff	Subterranean	388 627	19	23	1.08	219	21	198	2.19%	9.59%
3	Low saltern microbes (Pond 11)	Hypersaline	268 206	196	359	20.69	1 425	83	1 342	14.25%	5.82%
4	Medium saltern microbes (MB1110)	Hypersaline	38 929	43	71	29.28	1 882	130	1 752	18.82%	6.91%
5	Medium saltern microbes (MB1111)	Hypersaline	23 261	37	48	36.54	1 938	118	1 820	19.38%	6.09%
6	Low saltern pond plasmids (TT)	Hypersaline	111 431	117	179	26.56	1 205	70	1 135	12.05%	5.81%
7	Medium saltern microbial (MB1128)	Hypersaline	8 062	5	12	21.09	508	33	475	6.30%	6.50%
8	High saltern microbial (HB1128)	Hypersaline	35 446	14	41	15.52	2 321	188	2 133	23.21%	8.10%
9	Salton Sea Bacteria 1	Hypersaline	178 407	47	80	7.12	716	55	661	7.16%	7.68%
10	Medium salinity microbial (MB1116)	Hypersaline	120 987	134	221	29.34	2 171	167	2 004	21.71%	7.69%
11	Low salinity microbial (LB1128)	Hypersaline	34 296	27	41	19.83	595	28	567	5.95%	4.71%
24	Line Islands Kingman Reef B2 bacteria	Marine	188 445	11	23	1.80	334	15	319	3.34%	4.49%
25	Line Islands Christmas Reef B3 bacteria	Marine	227 542	18	42	2.64	192	14	178	1.92%	7.29%
26	Line Islands Palmyra F8 Bacteria	Marine	289 723	35	102	4.73	604	49	555	6.04%	8.11%
27	Line Islands Tabuaeran B1 Bacteria	Marine	290 844	4	48	1.79	200	28	172	2.00%	14.00%
28	DMSP 1 (MAM.1)	Marine	54 848	84	88	31.36	1 513	84	1 429	15.13%	5.55%
29	DMSP 2 (MAM.2)	Marine	50 313	88	112	39.75	962	53	909	9.62%	5.51%
30	VAN 1 (MAM 3)	Marine	12 446	17	19	28.92	908	67	841	9.08%	7.38%
31	VAN 2 (MAM 4)	Marine	33 773	39	81	35.53	1 317	91	1 226	13.17%	6.91%
42	Tilapia pond microbes	Freshw ater	381 076	130	325	11.94	1 265	107	1 158	12.65%	8.46%
43	Healthy Tilapia pond microbes	Freshw ater	63 978	46	77	19.23	1 143	92	1 051	11.43%	8.05%
44	Healthy Prebead tank microbes	Freshw ater	44 094	25	51	17.24	1 368	107	1 261	13.68%	7.82%
45	Tpond microbe 3	Freshw ater	67 612	62	122	27.21	1 294	89	1 205	12.94%	6.88%
50	Porites compressa time zero bacteria	Coral	53 473	2	17	3.55	144	20	124	1.44%	13.89%
51	Porites compressa control treated bacteria	Coral	65 191	4	19	3.53	111	14	97	1.11%	12.61%
52	Porites compressa temperature treated bacteria	Coral	61 356	17	18	5.70	91	15	76	0.91%	16.48%
53	Porites compressa DOC treated microbes	Coral	62 959	9	23	5.08	82	13	69	0.82%	15.85%
54	Porites compressa pH treated microbes	Coral	67 994	21	31	7.65	108	14	94	1.08%	12.96%
55	Porites compressa nutrient treated microbes	Coral	65 008	13	46	9.08	200	27	173	2.00%	13.50%
56	Porites astreoides microbial extraction	Coral	316 279	348	952	41.10	232	13	219	2.32%	5.60%
63	Rios Mesquites Stromatolites bacteria	Microbialites	124 694	11	16	2.17	1 881	82	1 799	18.81%	4.36%
64	Highborne Cay stromatolite bacteria	Microbialites	257 573	7	2	0.35	102	52	50	1.02%	50.98%
65	Pozas Azule II stromatolite microbes	Microbialites	326 146	45	130	5.37	833	75	758	8.33%	9.00%
69	Healthy slime bacteria	Fish	66 066	53	165	33.00	2 442	167	2 275	24.42%	6.84%
70	Morbid slime bacteria	Fish	82 442	105	345	54.58	2 701	194	2 507	27.01%	7.18%
71	Healthy gut bacteria	Fish	51 498	48	144	37.28	3 752	181	3 571	37.52%	4.82%
72	Morbid gut bacteria	Fish	60 311	65	229	48.75	3 696	459	3 237	36.96%	12.42%
75	Pooled Planktonic	Cow rumen	236 830	292	502	33.53	1 596	155	1 441	15.96%	9.71%
76	80F6	Cow rumen	178 713	224	398	34.80	1 754	166	1 588	17.54%	9.46%
77	640F6	Cow rumen	264 849	345	554	33.94	1 374	134	1 240	13.74%	9.75%
78	710F6	Cow rumen	345 317	645	1179	52.82	3 272	282	2 990	32.72%	8.62%
79	Lean Mice	Mice	49 074	31	53	17.12	8 245	1 414	6 831	82.45%	17.15%
80	Obese Mice	Mice	35 053	28	33	17.40	7 956	1 613	6 343	79.56%	20.27%
81	Chicken Cecum B	Chicken cecum	237 940	390	678	44.89	2 090	294	1 796	20.90%	14.07%
82	Chicken Cecum A	Chicken cecum	294 682	453	824	43.33	3 248	443	2 805	32.48%	13.64%

Table S3 : **Table of virome-genome association** where more than 500 reads from the virome were found to be similar to a non-prophage region of the genome. For each virome-genome pair, the number of reads with a hit is indicated, as well as the evaluation of the scattering of the reads from a recruitment plot (X : near complete coverage of the genome, - : coverage restricted to specific regions, ? : too few hits to assess), the percentage and number of genes covered by virome reads, and the presence of a GTA cluster in the genome,. Virome are colored as in Fig 1. All recruitment plots are available on http://metavir-meb.univ-bpclermont.fr/Recruitment_plots/recruitment_plot_gallery.php

Genome	Virome	Biome	Hit number	Cover – Plot	% Gene	Nb Genes	GTA
vok	MET7	Freshwater	1744	-	1.02	978	-
ajs	SRR089804	Eukaryote	2055	-	0.07	4383	-
pub	36	Marine	1973	X	56.23	1389	-
amc	21	Marine	1085	X	19.27	4146	-
sal	32	Marine	12435	X	70.71	3260	X
mlo	32	Marine	2028	X	16.38	7333	-
sal	33	Marine	6335	X	60.40	3260	X
mlo	33	Marine	1592	X	14.17	7333	-
sal	34	Marine	91315	X	79.82	3260	X
mlo	34	Marine	11602	X	33.68	7333	-
nar.pNL1	34	Marine	5954	X	47.59	187	X
eli	34	Marine	5647	X	14.78	3059	X
nar	34	Marine	4203	X	15.88	4031	X
swi	34	Marine	3981	X	14.12	5455	-
mes	34	Marine	3715	X	27.80	4684	-
hdn	34	Marine	2186	X	27.06	3600	-
pub	37	Marine	34799	X	91.36	1389	-
hse	73	Eukaryote	7162	X	47.17	4804	-
hse	74	Eukaryote	10548	X	55.25	4804	-
sbp	85	Eukaryote	19568	X	69.48	4666	-
sbn	85	Eukaryote	10004	X	52.83	4859	-
shw	85	Eukaryote	5748	X	40.60	4217	-
sbm	85	Eukaryote	5730	X	39.06	4596	-
son	85	Eukaryote	1821	?	9.80	4745	-
spe	86	Eukaryote	54182	X	65.50	5064	-
ssn	86	Eukaryote	15864	-	5.81	4804	-
pak	86	Eukaryote	12868	?	33.36	4956	-
sbc	86	Eukaryote	9600	-	0.39	5323	-
pau	86	Eukaryote	9345	-	21.67	5977	-
ecj	86	Eukaryote	9224	-	0.89	4494	-
sdy	86	Eukaryote	8047	-	0.80	4892	-
cpi	86	Eukaryote	6076	X	15.62	7399	-
mpo	86	Eukaryote	5237	?	11.22	5546	X
ecr	86	Eukaryote	3371	-	0.24	4635	-
ypy	86	Eukaryote	3365	X	21.26	4313	-
mea	86	Eukaryote	2559	-	7.63	6345	X
sty.pHCM1	86	Eukaryote	2539	-	2.85	246	-
yen	86	Eukaryote	2227	?	17.57	4280	-
pap	86	Eukaryote	1909	-	9.73	6369	-
swi	86	Eukaryote	1762	-	6.58	5455	-
pac	86	Eukaryote	1675	-	16.09	2368	-
mdi	86	Eukaryote	1665	-	4.46	6030	X
dia	86	Eukaryote	1660	?	7.20	3609	-
sjp.1	86	Eukaryote	1530	?	8.63	4461	X
pak	87	Eukaryote	22945	?	52.66	2410	-
pfs	87	Eukaryote	18589	X	61.80	6584	-
spe	87	Eukaryote	10150	X	45.70	5064	-
pac	87	Eukaryote	6797	-	30.32	2368	-
cpi	87	Eukaryote	1590	X	8.74	7399	-
bth	SRR043421	Eukaryote	3659	X	36.70	4902	-

Table S4 : **Detection of putative GTA clusters in KEGG genomes**. Each line describes a region considered as a putative GTA cluster (complete or fragmented), and displays the first and last gene of the region, the number of GTA-associated genes, number of viral-associated genes, and the taxonomic affiliation of the species involved.

gene_start	gene_stop	nb_genes	nb_GTA_genes	nb_viral_genes	species	kingdom	class	genus	species
rfe:RF_0311	rfe:RF_0318	8	1	2	rfe	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia felis
rfe:RF_0471	rfe:RF_0471	1	1	1	rfe	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia felis
rfe:RF_0479	rfe:RF_0479	1	1	1	rfe	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia felis
rfe:RF_0745	rfe:RF_0759	15	3	4	rfe	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia felis
rbe:RBE_0303	rbe:RBE_0311	9	3	4	rbe	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia bellii RML369-C
rbe:RBE_0629	rbe:RBE_0638	10	1	4	rbe	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia bellii RML369-C
rbe:RBE_0759	rbe:RBE_0759	1	1	1	rbe	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia bellii RML369-C
rbo:A1I_04355	rbo:A1I_04420	12	1	4	rbo	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia bellii OSU 85-389
rbo:A1I_04905	rbo:A1I_04905	1	1	1	rbo	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia bellii OSU 85-389
rbo:A1I_06250	rbo:A1I_06285	8	3	3	rbo	Bacteria	Alphaproteobacteria	Rickettsia	Rickettsia bellii OSU 85-389
ott:OTT_0461	ott:OTT_0468	8	1	3	ott	Bacteria	Alphaproteobacteria	Orientia	Orientia tsutsugamushi Ikeda
ott:OTT_0894	ott:OTT_0904	11	3	5	ott	Bacteria	Alphaproteobacteria	Orientia	Orientia tsutsugamushi Ikeda
ott:OTT_1321	ott:OTT_1321	1	1	1	ott	Bacteria	Alphaproteobacteria	Orientia	Orientia tsutsugamushi Ikeda
wol:WD0379	wol:WD0383	5	1	2	wol	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia wMel
wol:WD0443	wol:WD0449	6	1	2	wol	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia wMel
wol:WD0458	wol:WD0458	1	1	1	wol	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia wMel
wol:WD0744	wol:WD0749	7	1	1	wol	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia wMel
wol:WD1012	wol:WD1012	1	1	1	wol	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia wMel
wol:WD1311	wol:WD1311	1	1	1	wol	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia wMel
wri:WRi_002750	wri:WRi_002750	1	1	1	wri	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia sp. wRi
wri:WRi_002970	wri:WRi_003020	6	1	3	wri	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia sp. wRi
wri:WRi_004090	wri:WRi_004090	1	1	1	wri	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia sp. wRi
wri:WRi_007300	wri:WRi_007300	1	1	1	wri	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia sp. wRi
wri:WRi_009700	wri:WRi_009700	1	1	1	wri	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia sp. wRi
wri:WRi_013380	wri:WRi_013380	1	1	1	wri	Bacteria	Alphaproteobacteria	Wolbachia	Wolbachia sp. wRi
ama:AM397	ama:AM406	8	3	5	ama	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma marginale St. Maries
ama:AM809	ama:AM809	1	1	1	ama	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma marginale St. Maries
ama:AM1304	ama:AM1305	2	2	2	ama	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma marginale St. Maries
amf:AMF_293	amf:AMF_296	4	3	3	amf	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma marginale Florida
amf:AMF_601	amf:AMF_601	1	1	1	amf	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma marginale Florida
amf:AMF_985	amf:AMF_986	2	2	2	amf	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma marginale Florida
acn:ACIS_00100	acn:ACIS_00101	2	2	2	acn	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma centrale
acn:ACIS_00531	acn:ACIS_00531	1	1	1	acn	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma centrale
acn:ACIS_00894	acn:ACIS_00899	4	3	3	acn	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma centrale
aph:APH_0022	aph:APH_0023	2	2	2	aph	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma phagocytophilum
aph:APH_0091	aph:APH_0091	1	1	1	aph	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma phagocytophilum
aph:APH_0378	aph:APH_0378	1	1	1	aph	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma phagocytophilum
aph:APH_0680	aph:APH_0688	9	2	2	aph	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma phagocytophilum
aph:APH_0861	aph:APH_0861	1	1	1	aph	Bacteria	Alphaproteobacteria	Anaplasma	Anaplasma phagocytophilum
eru:Erum0200	eru:Erum0210	2	2	2	eru	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Welgevonden (South Africa)
eru:Erum2630	eru:Erum2660	4	3	3	eru	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Welgevonden (South Africa)
eru:Erum5190	eru:Erum5230	5	1	2	eru	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Welgevonden (South Africa)
erw:ERWE_CDS_00070	erw:ERWE_CDS_00080	2	2	2	erw	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Welgevonden (France)
erw:ERWE_CDS_02670	erw:ERWE_CDS_02700	4	3	3	erw	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Welgevonden (France)
erw:ERWE_CDS_05440	erw:ERWE_CDS_05500	7	1	2	erw	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Welgevonden (France)
erg:ERGA_CDS_00070	erg:ERGA_CDS_00080	2	2	2	erg	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Gardel
erg:ERGA_CDS_02630	erg:ERGA_CDS_02660	4	3	3	erg	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Gardel
erg:ERGA_CDS_05340	erg:ERGA_CDS_05390	6	1	2	erg	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia ruminantium Gardel
ecn:Ecaj_0012	ecn:Ecaj_0013	2	2	2	ecn	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia canis
ecn:Ecaj_0253	ecn:Ecaj_0256	4	3	3	ecn	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia canis
ecn:Ecaj_0528	ecn:Ecaj_0528	1	1	1	ecn	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia canis
ech:ECH_0032	ech:ECH_0033	2	2	2	ech	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia chaffeensis
ech:ECH_0500	ech:ECH_0500	1	1	1	ech	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia chaffeensis
ech:ECH_0830	ech:ECH_0830	1	1	1	ech	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia chaffeensis
ech:ECH_0835	ech:ECH_0836	2	2	2	ech	Bacteria	Alphaproteobacteria	Ehrlichia	Ehrlichia chaffeensis
pla:Plav_0902	pla:Plav_0923	22	12	13	pla	Bacteria	Alphaproteobacteria	Parvibaculum	Parvibaculum lavamentivorans
atu:Atu0948	atu:Atu8142	23	9	9	atu	Bacteria	Alphaproteobacteria	Agrobacterium	Agrobacterium tumefaciens C58
avi:Avi_1329	avi:Avi_1357	21	13	12	avi	Bacteria	Alphaproteobacteria	Agrobacterium	Agrobacterium vitis S4
bme:BMEI1338	bme:BMEI1350	13	9	7	bme	Bacteria	Alphaproteobacteria	Brucella	Brucella melitensis bv. 1 16M
bmi:BMEA_A0624	bmi:BMEA_A0641	18	10	9	bmi	Bacteria	Alphaproteobacteria	Brucella	Brucella melitensis ATCC 23457
bmf:BAB1_0608	bmf:BAB1_0627	20	10	9	bmf	Bacteria	Alphaproteobacteria	Brucella	Brucella melitensis biovar Abortus
bmb:BruAb1_0605	bmb:BruAb1_0622	18	9	9	bmb	Bacteria	Alphaproteobacteria	Brucella	Brucella abortus 9-941
bmc:BAbS19_I05690	bmc:BAbS19_I05870	18	11	10	bmc	Bacteria	Alphaproteobacteria	Brucella	Brucella abortus S19
bms:BR0584	bms:BR0603	20	11	10	bms	Bacteria	Alphaproteobacteria	Brucella	Brucella suis 1330
bmt:BSUIS_A0613	bmt:BSUIS_A0633	19	10	9	bmt	Bacteria	Alphaproteobacteria	Brucella	Brucella suis ATCC 23445
bov:BOV_0584	bov:BOV_0602	19	8	8	bov	Bacteria	Alphaproteobacteria	Brucella	Brucella ovis
bcs:BCAN_A0598	bcs:BCAN_A0618	20	11	10	bcs	Bacteria	Alphaproteobacteria	Brucella	Brucella canis
bmr:BMI_I583	bmr:BMI_I602	20	12	10	bmr	Bacteria	Alphaproteobacteria	Brucella	Brucella microti
rpa:RPA1885	rpa:RPA1908	23	11	8	rpa	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris CGA009
rpa:RPA1912	rpa:RPA1912	1	1	1	rpa	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris CGA009
rpa:RPA1914	rpa:RPA1914	1	1	1	rpa	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris CGA009
rpb:RPB_3455	rpb:RPB_3491	37	15	16	rpb	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris HaA2
rpc:RPC_1797	rpc:RPC_1829	33	12	10	rpc	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris BisB18
rpd:RPD_1964	rpd:RPD_1987	24	9	8	rpd	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris BisB5
rpd:RPD_1991	rpd:RPD_1996	6	4	5	rpd	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris BisB5
rpe:RPE_1893	rpe:RPE_1893	1	1	1	rpe	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris BisA53
rpe:RPE_1895	rpe:RPE_1898	4	2	2	rpe	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris BisA53
rpe:RPE_1902	rpe:RPE_1927	26	10	8	rpe	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris BisA53
rpt:Rpal_2092	rpt:Rpal_2102	11	4	5	rpt	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris TIE-1
rpt:Rpal_2106	rpt:Rpal_2116	11	7	4	rpt	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris TIE-1
rpt:Rpal_2122	rpt:Rpal_2122	1	1	1	rpt	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris TIE-1
rpt:Rpal_2124	rpt:Rpal_2124	1	1	1	rpt	Bacteria	Alphaproteobacteria	Rhodopseudomonas	Rhodopseudomonas palustris TIE-1

Table S4 .. continued :

gene_start	gene_stop	nb_genes	nb_GTA_genes	nb_viral_genes	species	kingdom	class	genus	species
nwi:Nwi_1159	nwi:Nwi_1186	28	12	11	nwi	Bacteria	Alphaproteobacteria	Nitrobacter	Nitrobacter winogradskyi
nha:Nham_1408	nha:Nham_1415	8	4	4	nha	Bacteria	Alphaproteobacteria	Nitrobacter	Nitrobacter hamburgensis
nha:Nham_1419	nha:Nham_1432	14	7	7	nha	Bacteria	Alphaproteobacteria	Nitrobacter	Nitrobacter hamburgensis
oca:OCAR_6586	oca:OCAR_6597	12	7	6	oca	Bacteria	Alphaproteobacteria	Oligotropha	Oligotropha carboxidovorans
oca:OCAR_6601	oca:OCAR_6611	11	4	4	oca	Bacteria	Alphaproteobacteria	Oligotropha	Oligotropha carboxidovorans
xau:Xaut_3144	xau:Xaut_3150	7	2	3	xau	Bacteria	Alphaproteobacteria	Xanthobacter	Xanthobacter autotrophicus
xau:Xaut_3154	xau:Xaut_3164	11	5	4	xau	Bacteria	Alphaproteobacteria	Xanthobacter	Xanthobacter autotrophicus
xau:Xaut_3168	xau:Xaut_3175	8	4	6	xau	Bacteria	Alphaproteobacteria	Xanthobacter	Xanthobacter autotrophicus
azc:AZC_1105	azc:AZC_1105	1	1	1	azc	Bacteria	Alphaproteobacteria	Azorhizobium	Azorhizobium caulinodans
azc:AZC_1109	azc:AZC_1128	20	10	11	azc	Bacteria	Alphaproteobacteria	Azorhizobium	Azorhizobium caulinodans
azc:AZC_3029	azc:AZC_3029	1	1	1	azc	Bacteria	Alphaproteobacteria	Azorhizobium	Azorhizobium caulinodans
sno:Snov_0699	sno:Snov_0724	26	11	13	sno	Bacteria	Alphaproteobacteria	Starkeya	Starkeya novella
mex:Mext_1412	mex:Mext_1434	23	11	12	mex	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium extorquens
mea:Mex_1p1299	mea:Mex_1p1322	24	11	12	mea	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium extorquens AM1
mdi:METDI2070	mdi:METDI2094	25	11	12	mdi	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium extorquens DM4
mrd:Mrad2831_4959	mrd:Mrad2831_4965	7	4	4	mrd	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium radiotolerans
mrd:Mrad2831_4969	mrd:Mrad2831_4981	13	8	5	mrd	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium radiotolerans
mpo:Mpop_1408	mpo:Mpop_1421	14	6	7	mpo	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium populi
mpo:Mpop_1426	mpo:Mpop_1431	6	4	5	mpo	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium populi
mch:Mchl_1686	mch:Mchl_1709	24	11	12	mch	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium chloromethanicum
mno:Mnod_2003	mno:Mnod_2018	16	4	7	mno	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium nodulans
mno:Mnod_2593	mno:Mnod_2618	26	5	19	mno	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium nodulans
mno:Mnod_4275	mno:Mnod_4299	25	5	20	mno	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium nodulans
mno:Mnod_6557	mno:Mnod_6577	21	4	12	mno	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium nodulans
mno:Mnod_7308	mno:Mnod_7342	35	6	13	mno	Bacteria	Alphaproteobacteria	Methylobacterium	Methylobacterium nodulans
msl:Msil_2179	msl:Msil_2179	1	1	1	msl	Bacteria	Alphaproteobacteria	Methylocella	Methylocella silvestris
msl:Msil_3038	msl:Msil_3069	32	4	8	msl	Bacteria	Alphaproteobacteria	Methylocella	Methylocella silvestris
ccr:CC_2772	ccr:CC_2790	19	10	10	ccr	Bacteria	Alphaproteobacteria	Caulobacter	Caulobacter crescentus CB15
ccs:CCNA_02859	ccs:CCNA_02877	18	9	9	ccs	Bacteria	Alphaproteobacteria	Caulobacter	Caulobacter crescentus NA1000
ccs:CCNA_02880	ccs:CCNA_02880	1	1	1	ccs	Bacteria	Alphaproteobacteria	Caulobacter	Caulobacter crescentus NA1000
cak:Caul_3900	cak:Caul_3920	21	9	9	cak	Bacteria	Alphaproteobacteria	Caulobacter	Caulobacter sp. K31
cse:Cseg_0983	cse:Cseg_1002	20	9	9	cse	Bacteria	Alphaproteobacteria	Caulobacter	Caulobacter segnis
bsb:Bresu_1222	bsb:Bresu_1249	28	3	10	bsb	Bacteria	Alphaproteobacteria	Brevundimonas	Brevundimonas subvibrioides
bsb:Bresu_1255	bsb:Bresu_1273	19	6	6	bsb	Bacteria	Alphaproteobacteria	Brevundimonas	Brevundimonas subvibrioides
sil:SPO2249	sil:SPO2267	17	17	9	sil	Bacteria	Alphaproteobacteria	Ruegeria	Silicibacter pomeroyi
sit:TM1040_1056	sit:TM1040_1074	19	18	10	sit	Bacteria	Alphaproteobacteria	Ruegeria	Ruegeria sp. TM1040
rsp:RSP_2465	rsp:RSP_2480	18	16	9	rsp	Bacteria	Alphaproteobacteria	Rhodobacter	Rhodobacter sphaeroides 2.4.1
rsh:Rsph17029_1128	rsh:Rsph17029_1145	18	17	9	rsh	Bacteria	Alphaproteobacteria	Rhodobacter	Rhodobacter sphaeroides ATCC 17029
rsq:Rsph17025_1073	rsq:Rsph17025_1090	18	17	9	rsq	Bacteria	Alphaproteobacteria	Rhodobacter	Rhodobacter sphaeroides ATCC 17025
rsk:RSKD131_0781	rsk:RSKD131_0797	17	17	9	rsk	Bacteria	Alphaproteobacteria	Rhodobacter	Rhodobacter sphaeroides KD131
rcp:RCAP_rcc01682	rcp:RCAP_rcc01699	18	18	9	rcp	Bacteria	Alphaproteobacteria	Rhodobacter	Rhodobacter capsulatus
jan:Jann_1631	jan:Jann_1647	17	16	11	jan	Bacteria	Alphaproteobacteria	Jannaschia	Jannaschia sp. CCS1
rde:RD1_3016	rde:RD1_3035	17	16	9	rde	Bacteria	Alphaproteobacteria	Roseobacter	Roseobacter denitrificans
pde:Pden_2886	pde:Pden_2902	17	16	10	pde	Bacteria	Alphaproteobacteria	Paracoccus	Paracoccus denitrificans
dsh:Dshi_2162	dsh:Dshi_2179	18	15	10	dsh	Bacteria	Alphaproteobacteria	Dinoroseobacter	Dinoroseobacter shibae
mmr:Mmar10_0917	mmr:Mmar10_0934	18	9	9	mmr	Bacteria	Alphaproteobacteria	Maricaulis	Maricaulis maris
hba:Hbal_1246	hba:Hbal_1255	10	4	4	hba	Bacteria	Alphaproteobacteria	Hirschia	Hirschia baltica
hba:Hbal_1259	hba:Hbal_1268	10	6	5	hba	Bacteria	Alphaproteobacteria	Hirschia	Hirschia baltica
nar:Saro_3116	nar:Saro_3133	18	8	8	nar	Bacteria	Alphaproteobacteria	Novosphingobium	Novosphingobium aromaticivorans
sal:Sala_1989	sal:Sala_2005	17	8	9	sal	Bacteria	Alphaproteobacteria	Sphingopyxis	Sphingopyxis alaskensis
sjp:SJA_C1-12210	sjp:SJA_C1-12390	19	8	7	sjp	Bacteria	Alphaproteobacteria	Sphingobium	Sphingobium japonicum
eli:ELI_13955	eli:ELI_14045	19	9	8	eli	Bacteria	Alphaproteobacteria	Erythrobacter	Erythrobacter litoralis
pbr:PB2503_08004	pbr:PB2503_08139	28	9	11	pbr	Bacteria	Alphaproteobacteria	Parvularcula	Parvularcula bermudensis
bhy:BHWA1_01823	bhy:BHWA1_01856	24	24	1	bhy	Bacteria	Spirochaetes	Brachyspira	Brachyspira hyodysenteriae
brm:Bmur_1423	brm:Bmur_1441	19	13	1	brm	Bacteria	Spirochaetes	Brachyspira	Brachyspira murdochii
bpo:BP951000_1199	bpo:BP951000_1209	11	11	0	bpo	Bacteria	Spirochaetes	Brachyspira	Brachyspira pilosicoli
bpo:BP951000_1214	bpo:BP951000_1221	8	5	2	bpo	Bacteria	Spirochaetes	Brachyspira	Brachyspira pilosicoli
mfe:Mefer_1536	mfe:Mefer_1553	19	7	3	mfe	Archaea	Euryarchaeota	Methanocaldococcus	Methanocaldococcus fervens
mvo:Mvol_0401	mvo:Mvol_0415	15	15	4	mvo	Archaea	Euryarchaeota	Methanococcus	Methanococcus voltae

Next pages ;

Table S5 : **Complete table of Kegg Orthologous groups retrieved in the viromes considered as free from any cellular DNA** (viromes with no rDNA detected). Groups are gathered by pathway, ordered by ko number.

Pathway	KO	Name	Definition	Class	Nb Reads	Nb Viromes
ko00010_Glycolysis / Gluconeogenesis						
	K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
	K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
	K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
	K00134	GAPDH, gapA	glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]	Carbohydrate Metabolism - Neurodegenerative Diseases	3	2
	K00161	PDHA, pdhA	pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	57	3
	K00162	PDHB, pdhB	pyruvate dehydrogenase E1 component subunit beta [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	37	3
	K00382	DLD, lpd, pdhD	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
	K00627	DLAT, aceF, pdhC	pyruvate dehydrogenase E2 component (dihydrolipoamide acetyltransferase) [EC:2.3.1.12]	Carbohydrate Metabolism	1	1
	K00845	glk	glucokinase [EC:2.7.1.2]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	2	2
	K00927	PGK, pgk	phosphoglycerate kinase [EC:2.7.2.3]	Carbohydrate Metabolism - Energy Metabolism	2	1
	K01610	E4.1.1.49, pckA	phosphoenolpyruvate carboxykinase (ATP) [EC:4.1.1.49]	Carbohydrate Metabolism - Energy Metabolism	4	2
	K01623	ALDO, fbaB	fructose-bisphosphate aldolase, class I [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	12	1
	K01624	FBA, fbaA	fructose-bisphosphate aldolase, class II [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	2	1
	K01792	E5.1.3.15	glucose-6-phosphate 1-epimerase [EC:5.1.3.15]	Carbohydrate Metabolism	1	1
	K01810	GPI, pgi	glucose-6-phosphate isomerase [EC:5.3.1.9]	Carbohydrate Metabolism	2	2
	K01895	ACSS, acs	acetyl-CoA synthetase [EC:6.2.1.1]	Carbohydrate Metabolism - Lipid Metabolism - Energy Metabolism	1	1
	K02791	PTS-Mal-EIIC, malX	PTS system, maltose and glucose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
	K04041	fbp3	fructose-1,6-bisphosphatase III [EC:3.1.3.11]	Carbohydrate Metabolism - Energy Metabolism	1	1
	K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
ko00020_Citrate cycle (TCA cycle)						
	K00024	mdh	malate dehydrogenase [EC:1.1.1.37]	Carbohydrate Metabolism - Energy Metabolism	3	3
	K00161	PDHA, pdhA	pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	57	3
	K00162	PDHB, pdhB	pyruvate dehydrogenase E1 component subunit beta [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	37	3
	K00164	OGDH, sucA	2-oxoglutarate dehydrogenase E1 component [EC:1.2.4.2]	Carbohydrate Metabolism - Amino Acid Metabolism	2	1
	K00174	korA	2-oxoglutarate ferredoxin oxidoreductase subunit alpha [EC:1.2.7.3]	Carbohydrate Metabolism - Energy Metabolism	1	1
	K00175	korB	2-oxoglutarate ferredoxin oxidoreductase subunit beta [EC:1.2.7.3]	Carbohydrate Metabolism - Energy Metabolism	2	2
	K00239	sdhA	succinate dehydrogenase flavoprotein subunit [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	2	2
	K00240	sdhB	succinate dehydrogenase iron-sulfur protein [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	18	1
	K00382	DLD, lpd, pdhD	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
	K00627	DLAT, aceF, pdhC	pyruvate dehydrogenase E2 component (dihydrolipoamide acetyltransferase) [EC:2.3.1.12]	Carbohydrate Metabolism	1	1
	K00658	DLST, sucB	2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61]	Carbohydrate Metabolism - Amino Acid Metabolism	2	2
	K01610	E4.1.1.49, pckA	phosphoenolpyruvate carboxykinase (ATP) [EC:4.1.1.49]	Carbohydrate Metabolism - Energy Metabolism	4	2
	K01681	ACO, acnA	aconitate hydratase 1 [EC:4.2.1.3]	Carbohydrate Metabolism - Energy Metabolism	4	3
ko00030_Pentose phosphate pathway						
	K00033	E1.1.1.44, PGD, gnd	6-phosphogluconate dehydrogenase [EC:1.1.1.44]	Carbohydrate Metabolism - Metabolism of Other Amino Acids	20	3
	K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49]	Carbohydrate Metabolism - Metabolism of Other Amino Acids	22	2
	K00615	E2.2.1.1, tktA, tktB	transketolase [EC:2.2.1.1]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Energy Metabolism	59	3
	K00616	E2.2.1.2, talA, talB	transaldolase [EC:2.2.1.2]	Carbohydrate Metabolism	46	4
	K00852	E2.7.1.15, rbsK	ribokinase [EC:2.7.1.15]	Carbohydrate Metabolism	3	2
	K00874	kdgK	2-dehydro-3-deoxygluconokinase [EC:2.7.1.45]	Carbohydrate Metabolism	1	1
	K00948	PRPS, prsA	ribose-phosphate pyrophosphokinase [EC:2.7.6.1]	Carbohydrate Metabolism - Nucleotide Metabolism	9	6
	K01619	E4.1.2.4, deoC	deoxyribose-phosphate aldolase [EC:4.1.2.4]	Carbohydrate Metabolism	3	2
	K01623	ALDO, fbaB	fructose-bisphosphate aldolase, class I [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	12	1
	K01624	FBA, fbaA	fructose-bisphosphate aldolase, class II [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	2	1
	K01783	rpe, RPE	ribulose-phosphate 3-epimerase [EC:5.1.3.1]	Carbohydrate Metabolism - Energy Metabolism	1	1
	K01808	E5.3.1.6B, rpiB	ribose 5-phosphate isomerase B [EC:5.3.1.6]	Carbohydrate Metabolism - Energy Metabolism	10	3
	K01810	GPI, pgi	glucose-6-phosphate isomerase [EC:5.3.1.9]	Carbohydrate Metabolism	2	2
	K04041	fbp3	fructose-1,6-bisphosphatase III [EC:3.1.3.11]	Carbohydrate Metabolism - Energy Metabolism	1	1
ko00040_Pentose and glucuronate interconversions						
	K00012	UGDH, ugd	UDPglucose 6-dehydrogenase [EC:1.1.1.22]	Carbohydrate Metabolism	91	5
	K00041	uxaB	tagaturonate reductase [EC:1.1.1.58]	Carbohydrate Metabolism	1	1
	K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
	K00848	rhaB	rhamnulokinase [EC:2.7.1.5]	Carbohydrate Metabolism	1	1
	K00874	kdgK	2-dehydro-3-deoxygluconokinase [EC:2.7.1.45]	Carbohydrate Metabolism	1	1
	K01783	rpe, RPE	ribulose-phosphate 3-epimerase [EC:5.1.3.1]	Carbohydrate Metabolism - Energy Metabolism	1	1
	K01786	araD	L-ribulose-5-phosphate 4-epimerase [EC:5.1.3.4]	Carbohydrate Metabolism	1	1
	K01812	uxaC	glucuronate isomerase [EC:5.3.1.12]	Carbohydrate Metabolism	1	1
	K03082	sgbU	hexulose-6-phosphate isomerase [EC:5.-.-.-]	Carbohydrate Metabolism	1	1
ko00051_Fructose and mannose metabolism						
	K00008	E1.1.1.14, gutB	L-iditol 2-dehydrogenase [EC:1.1.1.14]	Carbohydrate Metabolism	3	2
	K00066	algD	GDP-mannose 6-dehydrogenase [EC:1.1.1.132]	Carbohydrate Metabolism - Signal Transduction	3	1
	K00100	E1.1.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Lipid Metabolism	1	1
	K00754	E2.4.1.-		Carbohydrate Metabolism	149	4
	K00847	E2.7.1.4, scrK	fructokinase [EC:2.7.1.4]	Carbohydrate Metabolism	1	1
	K00848	rhaB	rhamnulokinase [EC:2.7.1.5]	Carbohydrate Metabolism	1	1
	K00966	GMPP	mannose-1-phosphate guanylyltransferase [EC:2.7.7.13]	Carbohydrate Metabolism	4	2
	K00971	E2.7.7.22, manC	mannose-1-phosphate guanylyltransferase [EC:2.7.7.22]	Carbohydrate Metabolism	12	4
	K01623	ALDO, fbaB	fructose-bisphosphate aldolase, class I [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	12	1
	K01624	FBA, fbaA	fructose-bisphosphate aldolase, class II [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	2	1
	K01628	fucA	L-fuculose-phosphate aldolase [EC:4.1.2.17]	Carbohydrate Metabolism	1	1
	K01711	E4.2.1.47, gmd	GDPmannose 4,6-dehydratase [EC:4.2.1.47]	Carbohydrate Metabolism	533	7
	K01809	E5.3.1.8, manA	mannose-6-phosphate isomerase [EC:5.3.1.8]	Carbohydrate Metabolism	17	3
	K01840	E5.4.2.8, manB	phosphomannomutase [EC:5.4.2.8]	Carbohydrate Metabolism	1	1
	K02377	E1.1.1.271, fcl	GDP-L-fucose synthase [EC:1.1.1.271]	Carbohydrate Metabolism	251	7
	K02770	PTS-Fru-EIIC, fruA	PTS system, fructose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
	K02795	PTS-Man-EIIC, manY	PTS system, mannose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
	K04041	fbp3	fructose-1,6-bisphosphatase III [EC:3.1.3.11]	Carbohydrate Metabolism - Energy Metabolism	1	1
ko00052_Galactose metabolism						
	K00845	glk	glucokinase [EC:2.7.1.2]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	2	2
	K01187	E3.2.1.20, malZ	alpha-glucosidase [EC:3.2.1.20]	Carbohydrate Metabolism	1	1
	K01190	lacZ	beta-galactosidase [EC:3.2.1.23]	Carbohydrate Metabolism - Glycan Biosynthesis and Metabolism - Lipid Metabolism	1	1
	K01193	E3.2.1.26, sacA	beta-fructofuranosidase [EC:3.2.1.26]	Carbohydrate Metabolism	1	1
	K01684	E4.2.1.6, dgoAb	galactonate dehydratase [EC:4.2.1.6]	Carbohydrate Metabolism	2	2
	K01784	galE, GALE	UDP-glucose 4-epimerase [EC:5.1.3.2]	Carbohydrate Metabolism	175	8
ko00053_Ascorbate and aldarate metabolism						
	K00012	UGDH, ugd	UDPglucose 6-dehydrogenase [EC:1.1.1.22]	Carbohydrate Metabolism	91	5

K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K03077	ulaF, sgaE	L-ribulose-5-phosphate 4-epimerase [EC:5.1.3.4]	Carbohydrate Metabolism	1	1
ko00061_Fatty acid biosynthesis					
K00059	fabG	3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]	Lipid Metabolism	10	4
K00647	fabB	3-oxoacyl-[acyl-carrier-protein] synthase I [EC:2.3.1.41]	Lipid Metabolism	6	2
K00648	fabH	3-oxoacyl-[acyl-carrier-protein] synthase III [EC:2.3.1.180]	Lipid Metabolism	6	3
K01962	accA	acetyl-CoA carboxylase carboxyl transferase subunit alpha [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	1	1
K01963	accD	acetyl-CoA carboxylase carboxyl transferase subunit beta [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	2	2
K02372	fabZ	3R-hydroxymyristoyl ACP dehydrase [EC:4.2.1.-]	Lipid Metabolism	1	1
K09458	fabF	3-oxoacyl-[acyl-carrier-protein] synthase II [EC:2.3.1.179]	Lipid Metabolism	4	2
K11263	bccA	acetyl-/propionyl-CoA carboxylase, biotin carboxylase, biotin carboxyl carrier protein [EC:6.3.4.14]	Lipid Metabolism	1	1
ko00071_Fatty acid metabolism					
K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00249	E1.3.99.3, ACADM, acd	acyl-CoA dehydrogenase [EC:1.3.99.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Endocrine System - Amino Acid Metabolism - Lipid Metabolism	1	1
K00529	hcaD	ferredoxin--NAD+ reductase [EC:1.18.1.3]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	Endocrine System - Transport and Catabolism - Lipid Metabolism	4	2
K05297	E1.18.1.1	rubredoxin-NAD+ reductase [EC:1.18.1.1]	Lipid Metabolism	1	1
K06445	fadE	acyl-CoA dehydrogenase [EC:1.3.99.-]	Lipid Metabolism	1	1
K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
ko00072_Synthesis and degradation of ketone bodies					
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K01640	E4.1.3.4, HMGCL, hmgL	hydroxymethylglutaryl-CoA lyase [EC:4.1.3.4]	Carbohydrate Metabolism - Transport and Catabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	18	1
ko00130_Ubiquinone and other terpenoid-quinone biosynthesis					
K03179	ubiA	4-hydroxybenzoate octaprenyltransferase [EC:2.5.1.-]	Metabolism of Cofactors and Vitamins - Metabolism of Terpenoids and Polyketides	1	1
K03183	ubiE	ubiquinone/menaquinone biosynthesis methyltransferase [EC:2.1.1.163 2.1.1.-]	Metabolism of Cofactors and Vitamins	2	2
ko00190_Oxidative phosphorylation					
K00239	sdhA	succinate dehydrogenase flavoprotein subunit [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	2	2
K00240	sdhB	succinate dehydrogenase iron-sulfur protein [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	18	1
K00331	nuoB	NADH dehydrogenase I subunit B [EC:1.6.5.3]	Energy Metabolism	2	2
K00336	nuoG	NADH dehydrogenase I subunit G [EC:1.6.5.3]	Energy Metabolism	1	1
K00337	nuoH	NADH dehydrogenase I subunit H [EC:1.6.5.3]	Energy Metabolism	2	2
K00341	nuoL	NADH dehydrogenase I subunit L [EC:1.6.5.3]	Energy Metabolism	2	2
K00342	nuoM	NADH dehydrogenase I subunit M [EC:1.6.5.3]	Energy Metabolism	1	1
K00343	nuoN	NADH dehydrogenase I subunit N [EC:1.6.5.3]	Energy Metabolism	1	1
K00405	ccoO	cb-type cytochrome c oxidase subunit II [EC:1.9.3.1]	Energy Metabolism	1	1
K00412	CYTB, petB	ubiquinol-cytochrome c reductase cytochrome b subunit [EC:1.10.2.2]	Circulatory System - Neurodegenerative Diseases - Energy Metabolism	7	2
K00425	cydA	cytochrome bd-I oxidase subunit I [EC:1.10.3.-]	Energy Metabolism	15	1
K00426	cydB	cytochrome bd-I oxidase subunit II [EC:1.10.3.-]	Energy Metabolism	2	2
K01507	E3.6.1.1, ppa	inorganic pyrophosphatase [EC:3.6.1.1]	Energy Metabolism	3	2
K02109	ATPF0B, atpF	F-type H+-transporting ATPase subunit b [EC:3.6.3.14]	Energy Metabolism	2	2
K02111	ATPF1A, atpA	F-type H+-transporting ATPase subunit alpha [EC:3.6.3.14]	Energy Metabolism	4	2
K02118	ATPVB, ntpB	V-type H+-transporting ATPase subunit B [EC:3.6.3.14]	Energy Metabolism	1	1
K02274	coxA	cytochrome c oxidase subunit I [EC:1.9.3.1]	Energy Metabolism	7	2
K02301	cyoE	protoheme IX farnesyltransferase [EC:2.5.1.-]	Metabolism of Cofactors and Vitamins - Metabolism of Terpenoids and Polyketides - Energy Metabolism	2	2
K05572	ndhA	NADH dehydrogenase I subunit 1 [EC:1.6.5.3]	Energy Metabolism	1	1
K05580	ndhI	NADH dehydrogenase I subunit I [EC:1.6.5.3]	Energy Metabolism	66	1
ko00195_Photosynthesis					
K02109	ATPF0B, atpF	F-type H+-transporting ATPase subunit b [EC:3.6.3.14]	Energy Metabolism	2	2
K02111	ATPF1A, atpA	F-type H+-transporting ATPase subunit alpha [EC:3.6.3.14]	Energy Metabolism	4	2
K02638	petE	plastocyanin	Energy Metabolism	5	2
K02639	petF	ferredoxin	Energy Metabolism	10	2
K02641	petH	ferredoxin--NADP+ reductase [EC:1.18.1.2]	Energy Metabolism	1	1
K02689	psaA	photosystem I P700 chlorophyll a apoprotein A1	Energy Metabolism	52	1
K02690	psaB	photosystem I P700 chlorophyll a apoprotein A2	Energy Metabolism	50	1
K02691	psaC	photosystem I subunit VII	Energy Metabolism	10	1
K02692	psaD	photosystem I subunit II	Energy Metabolism	10	1
K02693	psaE	photosystem I subunit IV	Energy Metabolism	2	1
K02694	psaF	photosystem I subunit III	Energy Metabolism	4	1
K02697	psaJ	photosystem I subunit IX	Energy Metabolism	2	1
K02703	psbA	photosystem II P680 reaction center D1 protein	Energy Metabolism	334	6
K02704	psbB	photosystem II CP47 chlorophyll apoprotein	Energy Metabolism	1	1
K02705	psbC	photosystem II CP43 chlorophyll apoprotein	Energy Metabolism	6	2
K02706	psbD	photosystem II P680 reaction center D2 protein	Energy Metabolism	129	4
ko00230_Purine metabolism					
K00088	E1.1.1.205, guaB	IMP dehydrogenase [EC:1.1.1.205]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	57	4
K00364	E1.7.1.7, guaC	GMP reductase [EC:1.7.1.7]	Nucleotide Metabolism	74	5
K00524	E1.17.4.1	ribonucleotide reductase, class II [EC:1.17.4.1]	Nucleotide Metabolism	3	2
K00525	E1.17.4.1A, nrdA, nrdE	ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]	Nucleotide Metabolism - Replication and Repair	3331	9
K00526	E1.17.4.1B, nrdB, nrdF	ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1]	Nucleotide Metabolism - Replication and Repair	869	8
K00527	nrdD	ribonucleoside-triphosphate reductase [EC:1.17.4.2]	Nucleotide Metabolism	57	7
K00759	E2.4.2.7, apt	adenine phosphoribosyltransferase [EC:2.4.2.7]	Nucleotide Metabolism	5	2
K00760	E2.4.2.8, hpt	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	62	4
K00764	E2.4.2.14, purF	amidophosphoribosyltransferase [EC:2.4.2.14]	Enzyme Families - Nucleotide Metabolism - Amino Acid Metabolism	2	2
K00860	cysC	adenylylsulfate kinase [EC:2.7.1.25]	Nucleotide Metabolism - Energy Metabolism	165	4
K00939	E2.7.4.3, adk	adenylate kinase [EC:2.7.4.3]	Nucleotide Metabolism	15	3
K00940	E2.7.4.6, ndk	nucleoside-diphosphate kinase [EC:2.7.4.6]	Nucleotide Metabolism	2	1
K00942	E2.7.4.8, gmk	guanylate kinase [EC:2.7.4.8]	Nucleotide Metabolism	1	1
K00948	PRPS, prsA	ribose-phosphate pyrophosphokinase [EC:2.7.6.1]	Carbohydrate Metabolism - Nucleotide Metabolism	9	6
K00951	relA	GTP pyrophosphokinase [EC:2.7.6.5]	Nucleotide Metabolism	3	2
K00955	cysNC	bifunctional enzyme CysN/CysC [EC:2.7.7.4 2.7.1.25]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Energy Metabolism	2	2

K00957	cysD	sulfate adenylyltransferase subunit 2 [EC:2.7.7.4]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Energy Metabolism	2	2
K00962	pnp, PNPT1	polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	Folding, Sorting and Degradation - Nucleotide Metabolism	1	1
K01081	E3.1.3.5	5'-nucleotidase [EC:3.1.3.5]	Signaling Molecules and Interaction - Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	3	3
K01428	ureC	urease subunit alpha [EC:3.5.1.5]	Infectious Diseases - Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	2	1
K01429	ureB	urease subunit beta [EC:3.5.1.5]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	1	1
K01486	ade	adenine deaminase [EC:3.5.4.2]	Nucleotide Metabolism	1	1
K01515	nudF	ADP-ribose pyrophosphatase [EC:3.6.1.13]	Nucleotide Metabolism	1	1
K01524	ppx-gppA	exopolyphosphatase / guanosine-5'-triphosphate,3'-diphosphate pyrophosphatase [EC:3.6.1.11 3.6.1.40]	Nucleotide Metabolism	1	1
K01588	purE	5-(carboxyamino)imidazole ribonucleotide mutase [EC:5.4.99.18]	Nucleotide Metabolism	1	1
K01756	E4.3.2.2, purB	adenylosuccinate lyase [EC:4.3.2.2]	Nucleotide Metabolism - Amino Acid Metabolism	3	3
K01768	E4.6.1.1	adenylate cyclase [EC:4.6.1.1]	Nucleotide Metabolism - Cell Growth and Death	61	5
K01923	purC	phosphoribosylaminoimidazole-succinocarboxamide synthase [EC:6.3.2.6]	Nucleotide Metabolism	1	1
K01933	purM	phosphoribosylformylglycinamidine cyclo-ligase [EC:6.3.3.1]	Nucleotide Metabolism	7	2
K01939	E6.3.4.4, purA	adenylosuccinate synthase [EC:6.3.4.4]	Nucleotide Metabolism - Amino Acid Metabolism	47	6
K01945	purD	phosphoribosylamine--glycine ligase [EC:6.3.4.13]	Nucleotide Metabolism	1	1
K01951	E6.3.5.2, guaA	GMP synthase (glutamine-hydrolysing) [EC:6.3.5.2]	Enzyme Families - Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	1	1
K01952	E6.3.5.3, purL	phosphoribosylformylglycinamidine synthase [EC:6.3.5.3]	Nucleotide Metabolism	1	1
K02319	DPA, polB1	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	468	5
K02322	DPB1	DNA polymerase II large subunit [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	41	2
K02323	DPB2	DNA polymerase II small subunit [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	5	2
K02335	DPO1, polA	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	921	8
K02337	DPO3A1, dnaE	DNA polymerase III subunit alpha [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	178	7
K02338	DPO3B, dnaN	DNA polymerase III subunit beta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	18	2
K02340	DPO3D1, holA	DNA polymerase III subunit delta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02341	DPO3D2, holB	DNA polymerase III subunit delta' [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02342	DPO3E, dnaQ	DNA polymerase III subunit epsilon [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	16	3
K02343	DPO3G, dnaX	DNA polymerase III subunit gamma/tau [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	13	3
K03040	rpoA	DNA-directed RNA polymerase subunit alpha [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	40	7
K03041	rpoA1	DNA-directed RNA polymerase subunit A' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	5	2
K03042	rpoA2	DNA-directed RNA polymerase subunit A" [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	6	1
K03043	rpoB	DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	5	5
K03044	rpoB1	DNA-directed RNA polymerase subunit B' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	4	3
K03045	rpoB2	DNA-directed RNA polymerase subunit B" [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	2	1
K03046	rpoC	DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	7	4
K03053	rpoH	DNA-directed RNA polymerase subunit H [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	1	1
K03056	rpoL	DNA-directed RNA polymerase subunit L [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	2	1
K03060	rpoZ	DNA-directed RNA polymerase subunit omega [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	1	1
K03763	DPO3A2, polC	DNA polymerase III subunit alpha, Gram-positive type [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	9	3
K03783	punA	purine-nucleoside phosphorylase [EC:2.4.2.1]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	1	1
K03784	deoD	purine-nucleoside phosphorylase [EC:2.4.2.1]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	2	2
K10807	RRM1	ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Replication and Repair	83	4
K10808	RRM2	ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Cell Growth and Death - Replication and Repair	67	5
K11175	purN	phosphoribosylglycinamide formyltransferase 1 [EC:2.1.2.2]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	12	2
K11177	yagR	xanthine dehydrogenase YagR molybdenum-binding subunit [EC:1.17.1.4]	Nucleotide Metabolism	1	1
K13798	K13798, rpoB	DNA-directed RNA polymerase subunit B [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	3	2

ko00240_Pyrimidine metabolism

K00226	pyrD	dihydroorotate oxidase [EC:1.3.3.1]	Nucleotide Metabolism	1	1
K00384	E1.8.1.9, trxB	thioredoxin reductase (NADPH) [EC:1.8.1.9]	Metabolism of Other Amino Acids - Nucleotide Metabolism	4	2
K00524	E1.17.4.1	ribonucleotide reductase, class II [EC:1.17.4.1]	Nucleotide Metabolism	3	2
K00525	E1.17.4.1A, nrdA, nrdE	ribonucleoside-diphosphate reductase alpha chain [EC:1.17.4.1]	Nucleotide Metabolism - Replication and Repair	3331	9
K00526	E1.17.4.1B, nrdB, nrdF	ribonucleoside-diphosphate reductase beta chain [EC:1.17.4.1]	Nucleotide Metabolism - Replication and Repair	869	8
K00527	nrdD	ribonucleoside-triphosphate reductase [EC:1.17.4.2]	Nucleotide Metabolism	57	7
K00560	E2.1.1.45, thyA	thymidylate synthase [EC:2.1.1.45]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	528	9
K00762	pyrE	orotate phosphoribosyltransferase [EC:2.4.2.10]	Nucleotide Metabolism	20	2
K00857	E2.7.1.21, tdk	thymidine kinase [EC:2.7.1.21]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	13	3
K00940	E2.7.4.6, ndk	nucleoside-diphosphate kinase [EC:2.7.4.6]	Nucleotide Metabolism	2	1
K00943	E2.7.4.9, tmk	dTMP kinase [EC:2.7.4.9]	Nucleotide Metabolism	64	2
K00945	cmk	cytidylate kinase [EC:2.7.4.14]	Nucleotide Metabolism	2	2
K00962	pnp, PNPT1	polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	Folding, Sorting and Degradation - Nucleotide Metabolism	1	1
K01081	E3.1.3.5	5'-nucleotidase [EC:3.1.3.5]	Signaling Molecules and Interaction - Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	3	3
K01465	URA4, pyrC	dihydroorotase [EC:3.5.2.3]	Nucleotide Metabolism	1	1
K01493	comEB	dCMP deaminase [EC:3.5.4.12]	Nucleotide Metabolism - Membrane Transport	145	8
K01494	E3.5.4.13, dcd	dCTP deaminase [EC:3.5.4.13]	Nucleotide Metabolism	588	5
K01520	E3.6.1.23, dut	dUTP pyrophosphatase [EC:3.6.1.23]	Nucleotide Metabolism - Replication and Repair	181	5
K01591	pyrF	orotidine-5'-phosphate decarboxylase [EC:4.1.1.23]	Nucleotide Metabolism	3	1
K01937	E6.3.4.2, pyrG	CTP synthase [EC:6.3.4.2]	Nucleotide Metabolism	1	1
K01955	carB, CPA2	carbamoyl-phosphate synthase large subunit [EC:6.3.5.5]	Nucleotide Metabolism - Amino Acid Metabolism	2	2
K01956	carA, CPA1	carbamoyl-phosphate synthase small subunit [EC:6.3.5.5]	Nucleotide Metabolism - Amino Acid Metabolism	2	1
K02319	DPA, polB1	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	468	5
K02322	DPB1	DNA polymerase II large subunit [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	41	2
K02323	DPB2	DNA polymerase II small subunit [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	5	2
K02335	DPO1, polA	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	921	8
K02337	DPO3A1, dnaE	DNA polymerase III subunit alpha [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	178	7
K02338	DPO3B, dnaN	DNA polymerase III subunit beta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	18	2
K02340	DPO3D1, holA	DNA polymerase III subunit delta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02341	DPO3D2, holB	DNA polymerase III subunit delta' [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02342	DPO3E, dnaQ	DNA polymerase III subunit epsilon [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	16	3
K02343	DPO3G, dnaX	DNA polymerase III subunit gamma/tau [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	13	3
K02825	pyrR	pyrimidine operon attenuation protein / uracil phosphoribosyltransferase [EC:2.4.2.9]	Transcription - Nucleotide Metabolism	1	1
K03040	rpoA	DNA-directed RNA polymerase subunit alpha [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	40	7
K03041	rpoA1	DNA-directed RNA polymerase subunit A' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	5	2
K03042	rpoA2	DNA-directed RNA polymerase subunit A" [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	6	1
K03043	rpoB	DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	5	5
K03044	rpoB1	DNA-directed RNA polymerase subunit B' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	4	3
K03045	rpoB2	DNA-directed RNA polymerase subunit B" [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	2	1
K03046	rpoC	DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	7	4
K03053	rpoH	DNA-directed RNA polymerase subunit H [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	1	1
K03056	rpoL	DNA-directed RNA polymerase subunit L [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	2	1
K03060	rpoZ	DNA-directed RNA polymerase subunit omega [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	1	1
K03465	E2.1.1.148, thyX, thy1	thymidylate synthase (FAD) [EC:2.1.1.148]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	1150	7
K03763	DPO3A2, polC	DNA polymerase III subunit alpha, Gram-positive type [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	9	3
K03783	punA	purine-nucleoside phosphorylase [EC:2.4.2.1]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	1	1
K03784	deoD	purine-nucleoside phosphorylase [EC:2.4.2.1]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	2	2
K09903	pyrH	uridylate kinase [EC:2.7.4.22]	Nucleotide Metabolism	1	1
K10807	RRM1	ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Replication and Repair	83	4
K10808	RRM2	ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Cell Growth and Death - Replication and Repair	67	5
K13421	UMPS	uridine monophosphate synthetase [EC:2.4.2.10 4.1.1.23]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	1	1
K13798	K13798, rpoB	DNA-directed RNA polymerase subunit B [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	3	2

ko00250_Alanine, aspartate and glutamate metabolism

K00135	E1.2.1.16, gabD	succinate-semialdehyde dehydrogenase (NADP+) [EC:1.2.1.16]	Carbohydrate Metabolism - Amino Acid Metabolism	2	1
K00261	E1.4.1.3	glutamate dehydrogenase (NAD(P)+) [EC:1.4.1.3]	Metabolism of Other Amino Acids - Excretory System - Amino Acid Metabolism - Energy Metabolism	3	2
K00265	gltB	glutamate synthase (NADPH/NADH) large chain [EC:1.4.1.13 1.4.1.14]	Amino Acid Metabolism - Energy Metabolism	3	3

K00266	gltD	glutamate synthase (NADPH/NADH) small chain [EC:1.4.1.13 1.4.1.14]	Amino Acid Metabolism - Energy Metabolism	1	1
K00278	nadB	L-aspartate oxidase [EC:1.4.3.16]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	1	1
K00294	E1.5.1.12	1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12]	Amino Acid Metabolism	1	1
K00764	E2.4.2.14, purF	amidophosphoribosyltransferase [EC:2.4.2.14]	Enzyme Families - Nucleotide Metabolism - Amino Acid Metabolism	2	2
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00820	E2.6.1.16, glmS	glucosamine--fructose-6-phosphate aminotransferase (isomerizing) [EC:2.6.1.16]	Carbohydrate Metabolism - Enzyme Families - Amino Acid Metabolism	20	4
K00823	puuE	4-aminobutyrate aminotransferase [EC:2.6.1.19]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K01424	E3.5.1.1, ansA, ansB	L-asparaginase [EC:3.5.1.1]	Metabolism of Other Amino Acids - Amino Acid Metabolism - Energy Metabolism	1	1
K01755	argH, ASL	argininosuccinate lyase [EC:4.3.2.1]	Amino Acid Metabolism	6	1
K01756	E4.3.2.2, purB	adenylosuccinate lyase [EC:4.3.2.2]	Nucleotide Metabolism - Amino Acid Metabolism	3	3
K01779	E5.1.1.13	aspartate racemase [EC:5.1.1.13]	Amino Acid Metabolism	1	1
K01915	E6.3.1.2, glnA	glutamine synthetase [EC:6.3.1.2]	Signal Transduction - Nervous System - Amino Acid Metabolism - Energy Metabolism	113	7
K01939	E6.3.4.4, purA	adenylosuccinate synthase [EC:6.3.4.4]	Nucleotide Metabolism - Amino Acid Metabolism	47	6
K01953	E6.3.5.4, asnB	asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4]	Enzyme Families - Amino Acid Metabolism - Energy Metabolism	14	2
K01955	carB, CPA2	carbamoyl-phosphate synthase large subunit [EC:6.3.5.5]	Nucleotide Metabolism - Amino Acid Metabolism	2	2
K01956	carA, CPA1	carbamoyl-phosphate synthase small subunit [EC:6.3.5.5]	Nucleotide Metabolism - Amino Acid Metabolism	2	1
K07250	gabT	4-aminobutyrate aminotransferase / (S)-3-amino-2-methylpropionate transaminase [EC:2.6.1.19 2.6.1.22]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K13821	putA	proline dehydrogenase / delta 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.99.8 1.5.1.12]	Amino Acid Metabolism	2	1

ko00253_Tetracycline biosynthesis

K01962	accA	acetyl-CoA carboxylase carboxyl transferase subunit alpha [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	1	1
K01963	accD	acetyl-CoA carboxylase carboxyl transferase subunit beta [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	2	2

ko00260_Glycine, serine and threonine metabolism

K00003	E1.1.1.3	homoserine dehydrogenase [EC:1.1.1.3]	Amino Acid Metabolism	1	1
K00050	E1.1.1.81, ttuD	hydroxypyruvate reductase [EC:1.1.1.81]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K00058	serA, PHGDH	D-3-phosphoglycerate dehydrogenase [EC:1.1.1.95]	Amino Acid Metabolism - Energy Metabolism	2	1
K00130	betB, gbsA	betaine-aldehyde dehydrogenase [EC:1.2.1.8]	Amino Acid Metabolism	1	1
K00273	E1.4.3.3, DAO	D-amino-acid oxidase [EC:1.4.3.3]	Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites - Transport and Catabolism - Amino Acid Metabolism	2	1
K00281	GLDC, gcvP	glycine dehydrogenase [EC:1.4.4.2]	Amino Acid Metabolism	10	3
K00282	gcvPA	glycine dehydrogenase subunit 1 [EC:1.4.4.2]	Amino Acid Metabolism	1	1
K00382	DLD, lpd, pdhD	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K00600	E2.1.2.1, glyA	glycine hydroxymethyltransferase [EC:2.1.2.1]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	25	4
K00605	E2.1.2.10, gcvT	aminomethyltransferase [EC:2.1.2.10]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	2	2
K00613	GATM	glycine amidinotransferase [EC:2.1.4.1]	Amino Acid Metabolism	5	3
K00643	E2.3.1.37, ALAS	5-aminolevulinate synthase [EC:2.3.1.37]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	7	4
K00831	serC, PSAT1	phosphoserine aminotransferase [EC:2.6.1.52]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	2	1
K00836	E2.6.1.76, ectB	diaminobutyrate-2-oxoglutarate transaminase [EC:2.6.1.76]	Amino Acid Metabolism	2	1
K00928	E2.7.2.4, lysC	aspartate kinase [EC:2.7.2.4]	Amino Acid Metabolism	1	1
K00998	E2.7.8.8, pssA	phosphatidylserine synthase [EC:2.7.8.8]	Amino Acid Metabolism - Lipid Metabolism	2	2
K01079	serB, PSPH	phosphoserine phosphatase [EC:3.1.3.3]	Amino Acid Metabolism - Energy Metabolism	4	3
K01620	E4.1.2.5, ltaA	threonine aldolase [EC:4.1.2.5]	Amino Acid Metabolism	1	1
K01696	E4.2.1.20B, trpB	tryptophan synthase beta chain [EC:4.2.1.20]	Amino Acid Metabolism	1	1
K01697	E4.2.1.22, CBS	cystathionine beta-synthase [EC:4.2.1.22]	Amino Acid Metabolism	1	1
K01754	E4.3.1.19, ilvA, tdcB	threonine dehydratase [EC:4.3.1.19]	Amino Acid Metabolism	3	2
K02203	thrH	phosphoserine / homoserine phosphotransferase [EC:3.1.3.3 2.7.1.39]	Amino Acid Metabolism - Energy Metabolism	1	1
K02204	E2.7.1.39B, thrB	homoserine kinase type II [EC:2.7.1.39]	Amino Acid Metabolism	1	1
K06720	ectC	L-ectoine synthase [EC:4.2.1.108]	Amino Acid Metabolism	19	1
K13745	ddc	L-2,4-diaminobutyrate decarboxylase [EC:4.1.1.86]	Amino Acid Metabolism	1	1

ko00270_Cysteine and methionine metabolism

K00003	E1.1.1.3	homoserine dehydrogenase [EC:1.1.1.3]	Amino Acid Metabolism	1	1
K00548	E2.1.1.13, metH	5-methyltetrahydrofolate--homocysteine methyltransferase [EC:2.1.1.13]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	23	5
K00558	E2.1.1.37, DNMT, dcm	DNA (cytosine-5-)-methyltransferase [EC:2.1.1.37]	Replication and Repair - Amino Acid Metabolism	2547	9
K00640	E2.3.1.30, cysE	serine O-acetyltransferase [EC:2.3.1.30]	Amino Acid Metabolism - Energy Metabolism	2	2
K00641	E2.3.1.31, metX	homoserine O-acetyltransferase [EC:2.3.1.31]	Amino Acid Metabolism - Energy Metabolism	1	1
K00789	E2.5.1.6, metK	S-adenosylmethionine synthetase [EC:2.5.1.6]	Amino Acid Metabolism	13	4
K00797	E2.5.1.16, SRM, speE	spermidine synthase [EC:2.5.1.16]	Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00928	E2.7.2.4, lysC	aspartate kinase [EC:2.7.2.4]	Amino Acid Metabolism	1	1
K01243	mtnN, mtn, pfs	S-adenosylhomocysteine/5'-methylthioadenosine nucleosidase [EC:3.2.2.9]	Amino Acid Metabolism	1	1
K01251	E3.3.1.1, ahcY	adenosylhomocysteinase [EC:3.3.1.1]	Amino Acid Metabolism	5	2
K01611	E4.1.1.50, speD	S-adenosylmethionine decarboxylase [EC:4.1.1.50]	Amino Acid Metabolism	100	5
K01697	E4.2.1.22, CBS	cystathionine beta-synthase [EC:4.2.1.22]	Amino Acid Metabolism	1	1
K01738	cysK	cysteine synthase A [EC:2.5.1.47]	Amino Acid Metabolism - Energy Metabolism	30	5
K01740	E2.5.1.49, metY	O-acetylhomoserine (thiol)-lyase [EC:2.5.1.49]	Amino Acid Metabolism	4	2
K01760	metC	cystathionine beta-lyase [EC:4.4.1.8]	Metabolism of Other Amino Acids - Amino Acid Metabolism - Energy Metabolism	2	1
K07173	luxS	S-ribosylhomocysteine lyase [EC:4.4.1.21]	Infectious Diseases - Amino Acid Metabolism	1	1
K08965	mtnW	2,3-diketo-5-methylthiopentyl-1-phosphate enolase [EC:3.1.3.77]	Amino Acid Metabolism	1	1
K10764	metZ	O-succinylhomoserine sulfhydrylase [EC:2.5.1.-]	Amino Acid Metabolism	1	1
K12339	cysM	cysteine synthase B [EC:2.5.1.47]	Amino Acid Metabolism - Energy Metabolism	5	2

ko00280_Valine, leucine and isoleucine degradation

K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00140	E1.2.1.27, mmsA, iolA	methylmalonate-semialdehyde dehydrogenase [EC:1.2.1.27]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K00167	E1.2.4.4B, bkdA2	2-oxoisovalerate dehydrogenase E1 component, beta subunit [EC:1.2.4.4]	Amino Acid Metabolism	1	1
K00249	E1.3.99.3, ACADM, acd	acyl-CoA dehydrogenase [EC:1.3.99.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Endocrine System - Amino Acid Metabolism - Lipid Metabolism	1	1
K00382	DLD, lpd, pdhD	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00826	E2.6.1.42, ilvE	branched-chain amino acid aminotransferase [EC:2.6.1.42]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	3	2
K01640	E4.1.3.4, HMGCL, hmgL	hydroxymethylglutaryl-CoA lyase [EC:4.1.3.4]	Carbohydrate Metabolism - Transport and Catabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	18	1
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01847	MUT	methylmalonyl-CoA mutase [EC:5.4.99.2]	Carbohydrate Metabolism - Amino Acid Metabolism - Energy Metabolism	2	2
K01848	E5.4.99.2A, mcmA1	methylmalonyl-CoA mutase, N-terminal domain [EC:5.4.99.2]	Carbohydrate Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K01969	E6.4.1.4B	3-methylcrotonyl-CoA carboxylase beta subunit [EC:6.4.1.4]	Amino Acid Metabolism	1	1
K07250	gabT	4-aminobutyrate aminotransferase / (S)-3-amino-2-methylpropionate transaminase [EC:2.6.1.19 2.6.1.22]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K11381	E1.2.4.4C, bkdA	2-oxoisovalerate dehydrogenase E1 component [EC:1.2.4.4]	Amino Acid Metabolism	6	2

ko00281_Geraniol degradation

K00257	E1.3.99.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides	2	2
K01640	E4.1.3.4, HMGCL, hmgL	hydroxymethylglutaryl-CoA lyase [EC:4.1.3.4]	Carbohydrate Metabolism - Transport and Catabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	18	1

K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1

ko00290_Valine, leucine and isoleucine biosynthesis

K00053	ilvC	ketol-acid reductoisomerase [EC:1.1.1.86]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	2	2
K00161	PDHA, pdhA	pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	57	3
K00162	PDHB, pdhB	pyruvate dehydrogenase E1 component subunit beta [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	37	3
K00826	E2.6.1.42, ilvE	branched-chain amino acid aminotransferase [EC:2.6.1.42]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	3	2
K01649	E2.3.3.13, leuA	2-isopropylmalate synthase [EC:2.3.3.13]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K01652	E2.2.1.6L, ilvB, ilvG, ilvI	acetolactate synthase I/II/III large subunit [EC:2.2.1.6]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	300	5
K01703	leuC	3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit [EC:4.2.1.33 4.2.1.35]	Carbohydrate Metabolism - Amino Acid Metabolism	2	2
K01704	leuD	3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit [EC:4.2.1.33 4.2.1.35]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K01754	E4.3.1.19, ilvA, tdcB	threonine dehydratase [EC:4.3.1.19]	Amino Acid Metabolism	3	2
K01869	LARS, leuS	leucyl-tRNA synthetase [EC:6.1.1.4]	Translation - Amino Acid Metabolism	2	2
K01870	IARS, ileS	isoleucyl-tRNA synthetase [EC:6.1.1.5]	Translation - Amino Acid Metabolism	3	2
K01873	VARS, valS	valyl-tRNA synthetase [EC:6.1.1.9]	Translation - Amino Acid Metabolism	3	2

ko00300_Lysine biosynthesis

K00003	E1.1.1.3	homoserine dehydrogenase [EC:1.1.1.3]	Amino Acid Metabolism	1	1
K00145	argC	N-acetyl-gamma-glutamyl-phosphate/N-acetyl-gamma-aminoadipyl-phosphate reductase [EC:1.2.1.38 1.2.1.-]	Amino Acid Metabolism	1	1
K00215	dapB	dihydrodipicolinate reductase [EC:1.3.1.26]	Amino Acid Metabolism	1	1
K00674	dapD	2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase [EC:2.3.1.117]	Amino Acid Metabolism	4	2
K00928	E2.7.2.4, lysC	aspartate kinase [EC:2.7.2.4]	Amino Acid Metabolism	1	1
K01439	dapE	succinyl-diaminopimelate desuccinylase [EC:3.5.1.18]	Amino Acid Metabolism	1	1
K01586	lysA	diaminopimelate decarboxylase [EC:4.1.1.20]	Amino Acid Metabolism	3	2
K01714	dapA	dihydrodipicolinate synthase [EC:4.2.1.52]	Amino Acid Metabolism	1	1
K01778	dapF	diaminopimelate epimerase [EC:5.1.1.7]	Amino Acid Metabolism	1	1
K01929	murF	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanine ligase [EC:6.3.2.10]	Glycan Biosynthesis and Metabolism - Amino Acid Metabolism	1	1
K03918	lat	L-lysine 6-transaminase [EC:2.6.1.36]	Amino Acid Metabolism	2	2

ko00310_Lysine degradation

K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00164	OGDH, sucA	2-oxoglutarate dehydrogenase E1 component [EC:1.2.4.2]	Carbohydrate Metabolism - Amino Acid Metabolism	2	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00658	DLST, sucB	2-oxoglutarate dehydrogenase E2 component (dihydrolipoamide succinyltransferase) [EC:2.3.1.61]	Carbohydrate Metabolism - Amino Acid Metabolism	2	2
K01423	E3.4.-.-		Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	7	4
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01843	E5.4.3.2, kamA	lysine 2,3-aminomutase [EC:5.4.3.2]	Amino Acid Metabolism	2	1

ko00311_Penicillin and cephalosporin biosynthesis

K00273	E1.4.3.3, DAO	D-amino-acid oxidase [EC:1.4.3.3]	Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites - Transport and Catabolism - Amino Acid Metabolism	2	1
K01434	E3.5.1.11	penicillin amidase [EC:3.5.1.11]	Enzyme Families - Biosynthesis of Other Secondary Metabolites	1	1
K01467	E3.5.2.6, ampC, penP	beta-lactamase [EC:3.5.2.6]	Signal Transduction - Biosynthesis of Other Secondary Metabolites	1	1

ko00312_beta-Lactam resistance

K01467	E3.5.2.6, ampC, penP	beta-lactamase [EC:3.5.2.6]	Signal Transduction - Biosynthesis of Other Secondary Metabolites	1	1

ko00330_Arginine and proline metabolism

K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00145	argC	N-acetyl-gamma-glutamyl-phosphate/N-acetyl-gamma-aminoadipyl-phosphate reductase [EC:1.2.1.38 1.2.1.-]	Amino Acid Metabolism	1	1
K00147	proA	glutamate-5-semialdehyde dehydrogenase [EC:1.2.1.41]	Amino Acid Metabolism	1	1
K00261	E1.4.1.3	glutamate dehydrogenase (NAD(P)+) [EC:1.4.1.3]	Metabolism of Other Amino Acids - Excretory System - Amino Acid Metabolism - Energy Metabolism	3	2
K00273	E1.4.3.3, DAO	D-amino-acid oxidase [EC:1.4.3.3]	Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites - Transport and Catabolism - Amino Acid Metabolism	2	1
K00294	E1.5.1.12	1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.1.12]	Amino Acid Metabolism	1	1
K00472	E1.14.11.2	prolyl 4-hydroxylase [EC:1.14.11.2]	Amino Acid Metabolism	25	4
K00613	GATM	glycine amidinotransferase [EC:2.1.4.1]	Amino Acid Metabolism	5	3
K00797	E2.5.1.16, SRM, speE	spermidine synthase [EC:2.5.1.16]	Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00931	proB	glutamate 5-kinase [EC:2.7.2.11]	Amino Acid Metabolism	2	2
K01426	E3.5.1.4, amiE	amidase [EC:3.5.1.4]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	3	1
K01428	ureC	urease subunit alpha [EC:3.5.1.5]	Infectious Diseases - Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	2	1
K01429	ureB	urease subunit beta [EC:3.5.1.5]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	1	1
K01438	E3.5.1.16, argE	acetylornithine deacetylase [EC:3.5.1.16]	Amino Acid Metabolism	1	1
K01476	E3.5.3.1, rocF, arg	arginase [EC:3.5.3.1]	Infectious Diseases - Amino Acid Metabolism	11	1
K01480	E3.5.3.11, speB	agmatinase [EC:3.5.3.11]	Amino Acid Metabolism	2	1
K01581	E4.1.1.17, ODC1, speC, speF	ornithine decarboxylase [EC:4.1.1.17]	Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K01611	E4.1.1.50, speD	S-adenosylmethionine decarboxylase [EC:4.1.1.50]	Amino Acid Metabolism	100	5
K01750	E4.3.1.12, ocd	ornithine cyclodeaminase [EC:4.3.1.12]	Amino Acid Metabolism	1	1
K01755	argH, ASL	argininosuccinate lyase [EC:4.3.2.1]	Amino Acid Metabolism	6	1
K01915	E6.3.1.2, glnA	glutamine synthetase [EC:6.3.1.2]	Signal Transduction - Nervous System - Amino Acid Metabolism - Energy Metabolism	113	7
K09251	E2.6.1.82	putrescine aminotransferase [EC:2.6.1.82]	Amino Acid Metabolism	2	1
K09472	puuC, aldH	gamma-glutamyl-gamma-aminobutyraldehyde dehydrogenase [EC:1.2.1.-]	Amino Acid Metabolism	1	1
K10536	E3.5.3.12	agmatine deiminase [EC:3.5.3.12]	Amino Acid Metabolism	2	1
K13747	nspC	carboxynorspermidine decarboxylase [EC:4.1.1.-]	Amino Acid Metabolism	1	1
K13821	putA	proline dehydrogenase / delta 1-pyrroline-5-carboxylate dehydrogenase [EC:1.5.99.8 1.5.1.12]	Amino Acid Metabolism	2	1

ko00340_Histidine metabolism

K00013	hisD	histidinol dehydrogenase [EC:1.1.1.23]	Amino Acid Metabolism	1	1
K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00599	E2.1.1.-		Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	55	3
K00817	hisC	histidinol-phosphate aminotransferase [EC:2.6.1.9]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	6	3
K01468	E3.5.2.7, hutI	imidazolonepropionase [EC:3.5.2.7]	Amino Acid Metabolism	1	1
K01693	E4.2.1.19, hisB	imidazoleglycerol-phosphate dehydratase [EC:4.2.1.19]	Amino Acid Metabolism	1	1

K01814	E5.3.1.16, hisA	phosphoribosylformimino-5-aminoimidazole carboxamide ribotide isomerase [EC:5.3.1.16]	Amino Acid Metabolism	1	1
K02500	hisF	cyclase [EC:4.1.3.-]	Amino Acid Metabolism	2	1
K02501	hisH	glutamine amidotransferase [EC:2.4.2.-]	Amino Acid Metabolism	1	1

ko00350_Tyrosine metabolism

K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K00135	E1.2.1.16, gabD	succinate-semialdehyde dehydrogenase (NADP+) [EC:1.2.1.16]	Carbohydrate Metabolism - Amino Acid Metabolism	2	1
K00450	E1.13.11.4	gentisate 1,2-dioxygenase [EC:1.13.11.4]	Amino Acid Metabolism	1	1
K00599	E2.1.1.-		Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	55	3
K00680	E2.3.1.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism	9	1
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00817	hisC	histidinol-phosphate aminotransferase [EC:2.6.1.9]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	6	3
K02510	hpaI	2,4-dihydroxyhept-2-ene-1,7-dioic acid aldolase [EC:4.1.2.-]	Amino Acid Metabolism	2	2
K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1

ko00360_Phenylalanine metabolism

K00285	dadA	D-amino-acid dehydrogenase [EC:1.4.99.1]	Amino Acid Metabolism - Energy Metabolism	19	4
K00529	hcaD	ferredoxin--NAD+ reductase [EC:1.18.1.3]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00588	E2.1.1.104	caffeoyl-CoA O-methyltransferase [EC:2.1.1.104]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	1	1
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00817	hisC	histidinol-phosphate aminotransferase [EC:2.6.1.9]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	6	3
K01426	E3.5.1.4, amiE	amidase [EC:3.5.1.4]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	3	1
K01666	E4.1.3.39, mhpE	4-hydroxy 2-oxovalerate aldolase [EC:4.1.3.39]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	35	3
K01912	E6.2.1.30, paaK	phenylacetate-CoA ligase [EC:6.2.1.30]	Amino Acid Metabolism	2	2
K02554	mhpD	2-keto-4-pentenoate hydratase [EC:4.2.1.80]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	1	1
K03782	katG	catalase/peroxidase [EC:1.11.1.6 1.11.1.7]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	9	1
K05709	hcaA2, hcaF	small terminal subunit of phenylpropionate dioxygenase [EC:1.14.12.19]	Amino Acid Metabolism	1	1

ko00361_Chlorocyclohexane and chlorobenzene degradation

K01563	dhaA	haloalkane dehalogenase [EC:3.8.1.5]	Xenobiotics Biodegradation and Metabolism	1	1

ko00362_Benzoate degradation

K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00680	E2.3.1.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism	9	1
K01666	E4.1.3.39, mhpE	4-hydroxy 2-oxovalerate aldolase [EC:4.1.3.39]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	35	3
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01726	E4.2.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides	1	1
K02554	mhpD	2-keto-4-pentenoate hydratase [EC:4.2.1.80]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	1	1
K04073	mhpF	acetaldehyde dehydrogenase [EC:1.2.1.10]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism	43	2
K05783	benD	1,6-dihydroxycyclohexa-2,4-diene-1-carboxylate dehydrogenase [EC:1.3.1.25]	Xenobiotics Biodegradation and Metabolism	1	1

ko00363_Bisphenol degradation

K00100	E1.1.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Lipid Metabolism	1	1
K01726	E4.2.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides	1	1

ko00364_Fluorobenzoate degradation

K05783	benD	1,6-dihydroxycyclohexa-2,4-diene-1-carboxylate dehydrogenase [EC:1.3.1.25]	Xenobiotics Biodegradation and Metabolism	1	1

ko00380_Tryptophan metabolism

K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00164	OGDH, sucA	2-oxoglutarate dehydrogenase E1 component [EC:1.2.4.2]	Carbohydrate Metabolism - Amino Acid Metabolism	2	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K01426	E3.5.1.4, amiE	amidase [EC:3.5.1.4]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	3	1
K01501	E3.5.5.1	nitrilase [EC:3.5.5.1]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K01556	E3.7.1.3	kynureninase [EC:3.7.1.3]	Amino Acid Metabolism	1	1
K01667	E4.1.99.1, tnaA	tryptophanase [EC:4.1.99.1]	Amino Acid Metabolism - Energy Metabolism	1	1
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01867	WARS, trpS	tryptophanyl-tRNA synthetase [EC:6.1.1.2]	Translation - Amino Acid Metabolism	3	1
K03781	katE, CAT	catalase [EC:1.11.1.6]	Neurodegenerative Diseases - Transport and Catabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K03782	katG	catalase/peroxidase [EC:1.11.1.6 1.11.1.7]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	9	1
K04103	E4.1.1.74, ipdC	indolepyruvate decarboxylase [EC:4.1.1.74]	Amino Acid Metabolism	1	1

ko00400_Phenylalanine, tyrosine and tryptophan biosynthesis

K00014	aroE	shikimate dehydrogenase [EC:1.1.1.25]	Amino Acid Metabolism	3	1
K00210	E1.3.1.12	prephenate dehydrogenase [EC:1.3.1.12]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	1	1
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00817	hisC	histidinol-phosphate aminotransferase [EC:2.6.1.9]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	6	3
K01657	trpE	anthranilate synthase component I [EC:4.1.3.27]	Amino Acid Metabolism	2	2
K01696	E4.2.1.20B, trpB	tryptophan synthase beta chain [EC:4.2.1.20]	Amino Acid Metabolism	1	1
K01735	aroB	3-dehydroquinate synthase [EC:4.2.3.4]	Amino Acid Metabolism	28	4
K03856	AROA2, aroA	3-deoxy-7-phosphoheptulonate synthase [EC:2.5.1.54]	Amino Acid Metabolism	3	3
K04516	AROA1, aroA	chorismate mutase [EC:5.4.99.5]	Amino Acid Metabolism	1	1
K04518	pheA2	prephenate dehydratase [EC:4.2.1.51]	Amino Acid Metabolism	1	1
K13853	aroG, aroA	3-deoxy-7-phosphoheptulonate synthase / chorismate mutase [EC:2.5.1.54 5.4.99.5]	Amino Acid Metabolism	1	1
K14170	pheA	chorismate mutase / prephenate dehydratase [EC:5.4.99.5 4.2.1.51]	Amino Acid Metabolism	1	1

ko00401_Novobiocin biosynthesis

K00210	E1.3.1.12	prephenate dehydrogenase [EC:1.3.1.12]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	1	1
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00817	hisC	histidinol-phosphate aminotransferase [EC:2.6.1.9]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	6	3

ko00410_beta-Alanine metabolism

K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4

K00249	E1.3.99.3, ACADM, acd	acyl-CoA dehydrogenase [EC:1.3.99.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Endocrine System - Amino Acid Metabolism - Lipid Metabolism	1	1
K00797	E2.5.1.16, SRM, speE	spermidine synthase [EC:2.5.1.16]	Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K00823	puuE	4-aminobutyrate aminotransferase [EC:2.6.1.19]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K07250	gabT	4-aminobutyrate aminotransferase / (S)-3-amino-2-methylpropionate transaminase [EC:2.6.1.19 2.6.1.22]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
ko00440_Phosphonate and phosphinate metabolism					
K00968	E2.7.7.15, PCYT1	choline-phosphate cytidylyltransferase [EC:2.7.7.15]	Metabolism of Other Amino Acids - Lipid Metabolism	12	3
K01841	E5.4.2.9	phosphoenolpyruvate phosphomutase [EC:5.4.2.9]	Metabolism of Other Amino Acids	2	1
ko00450_Selenocompound metabolism					
K00384	E1.8.1.9, trxB	thioredoxin reductase (NADPH) [EC:1.8.1.9]	Metabolism of Other Amino Acids - Nucleotide Metabolism	4	2
K00548	E2.1.1.13, metH	5-methyltetrahydrofolate--homocysteine methyltransferase [EC:2.1.1.13]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	23	5
K00955	cysNC	bifunctional enzyme CysN/CysC [EC:2.7.7.4 2.7.1.25]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Energy Metabolism	2	2
K00957	cysD	sulfate adenylyltransferase subunit 2 [EC:2.7.7.4]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Energy Metabolism	2	2
K01760	metC	cystathionine beta-lyase [EC:4.4.1.8]	Metabolism of Other Amino Acids - Amino Acid Metabolism - Energy Metabolism	2	1
K01874	MARS, metG	methionyl-tRNA synthetase [EC:6.1.1.10]	Translation - Metabolism of Other Amino Acids - Amino Acid Metabolism	6	3
K11717	sufS	cysteine desulfurase / selenocysteine lyase [EC:2.8.1.7 4.4.1.16]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins	1	1
ko00460_Cyanoamino acid metabolism					
K00600	E2.1.2.1, glyA	glycine hydroxymethyltransferase [EC:2.1.2.1]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	25	4
K01424	E3.5.1.1, ansA, ansB	L-asparaginase [EC:3.5.1.1]	Metabolism of Other Amino Acids - Amino Acid Metabolism - Energy Metabolism	1	1
K01426	E3.5.1.4, amiE	amidase [EC:3.5.1.4]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	3	1
K01501	E3.5.5.1	nitrilase [EC:3.5.5.1]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K05349	bglX	beta-glucosidase [EC:3.2.1.21]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites	1	1
K05350	bglB	beta-glucosidase [EC:3.2.1.21]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites	4	2
ko00471_D-Glutamine and D-glutamate metabolism					
K00261	E1.4.1.3	glutamate dehydrogenase (NAD(P)+) [EC:1.4.1.3]	Metabolism of Other Amino Acids - Excretory System - Amino Acid Metabolism - Energy Metabolism	3	2
K01925	murD	UDP-N-acetylmuramoylalanine--D-glutamate ligase [EC:6.3.2.9]	Metabolism of Other Amino Acids - Glycan Biosynthesis and Metabolism	2	2
ko00472_D-Arginine and D-ornithine metabolism					
K00273	E1.4.3.3, DAO	D-amino-acid oxidase [EC:1.4.3.3]	Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites - Transport and Catabolism - Amino Acid Metabolism	2	1
ko00473_D-Alanine metabolism					
K01921	ddl	D-alanine-D-alanine ligase [EC:6.3.2.4]	Metabolism of Other Amino Acids - Glycan Biosynthesis and Metabolism	2	2
ko00480_Glutathione metabolism					
K00033	E1.1.1.44, PGD, gnd	6-phosphogluconate dehydrogenase [EC:1.1.1.44]	Carbohydrate Metabolism - Metabolism of Other Amino Acids	20	3
K00036	G6PD, zwf	glucose-6-phosphate 1-dehydrogenase [EC:1.1.1.49]	Carbohydrate Metabolism - Metabolism of Other Amino Acids	22	2
K00432	E1.11.1.9	glutathione peroxidase [EC:1.11.1.9]	Metabolism of Other Amino Acids - Lipid Metabolism	2	2
K00797	E2.5.1.16, SRM, speE	spermidine synthase [EC:2.5.1.16]	Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K01255	CARP, pepA	leucyl aminopeptidase [EC:3.4.11.1]	Metabolism of Other Amino Acids - Enzyme Families	3	1
K01581	E4.1.1.17, ODC1, speC, speF	ornithine decarboxylase [EC:4.1.1.17]	Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K10807	RRM1	ribonucleoside-diphosphate reductase subunit M1 [EC:1.17.4.1]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Replication and Repair	83	4
K10808	RRM2	ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Cell Growth and Death - Replication and Repair	67	5
ko00500_Starch and sucrose metabolism					
K00012	UGDH, ugd	UDPglucose 6-dehydrogenase [EC:1.1.1.22]	Carbohydrate Metabolism	91	5
K00688	E2.4.1.1, glgP, PYG	starch phosphorylase [EC:2.4.1.1]	Carbohydrate Metabolism - Endocrine System	9	3
K00697	E2.4.1.15, otsA	alpha,alpha-trehalose-phosphate synthase (UDP-forming) [EC:2.4.1.15]	Carbohydrate Metabolism - Glycan Biosynthesis and Metabolism	3	1
K00705	malQ	4-alpha-glucanotransferase [EC:2.4.1.25]	Carbohydrate Metabolism	1	1
K00845	glk	glucokinase [EC:2.7.1.2]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	2	2
K00847	E2.7.1.4, scrK	fructokinase [EC:2.7.1.4]	Carbohydrate Metabolism	1	1
K00978	rfbF	glucose-1-phosphate cytidylyltransferase [EC:2.7.7.33]	Carbohydrate Metabolism	6	1
K01176	E3.2.1.1, amyA, malS	alpha-amylase [EC:3.2.1.1]	Carbohydrate Metabolism - Digestive System	2	1
K01179	E3.2.1.4	endoglucanase [EC:3.2.1.4]	Carbohydrate Metabolism	11	2
K01187	E3.2.1.20, malZ	alpha-glucosidase [EC:3.2.1.20]	Carbohydrate Metabolism	1	1
K01193	E3.2.1.26, sacA	beta-fructofuranosidase [EC:3.2.1.26]	Carbohydrate Metabolism	1	1
K01225	E3.2.1.91	cellulose 1,4-beta-cellobiosidase [EC:3.2.1.91]	Carbohydrate Metabolism	7	2
K01810	GPI, pgi	glucose-6-phosphate isomerase [EC:5.3.1.9]	Carbohydrate Metabolism	2	2
K02791	PTS-Mal-EIIC, malX	PTS system, maltose and glucose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
K05349	bglX	beta-glucosidase [EC:3.2.1.21]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites	1	1
K05350	bglB	beta-glucosidase [EC:3.2.1.21]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites	4	2
ko00510_N-Glycan biosynthesis					
K00721	DPM1	dolichol-phosphate mannosyltransferase [EC:2.4.1.83]	Glycan Biosynthesis and Metabolism	39	3
K00737	MGAT3	beta-1,4-mannosyl-glycoprotein beta-1,4-N-acetylglucosaminyltransferase [EC:2.4.1.144]	Glycan Biosynthesis and Metabolism	23	3
ko00511_Other glycan degradation					
K01190	lacZ	beta-galactosidase [EC:3.2.1.23]	Carbohydrate Metabolism - Glycan Biosynthesis and Metabolism - Lipid Metabolism	1	1
K01191	E3.2.1.24	alpha-mannosidase [EC:3.2.1.24]	Glycan Biosynthesis and Metabolism	2	1
K01206	E3.2.1.51, FUCA	alpha-L-fucosidase [EC:3.2.1.51]	Glycan Biosynthesis and Metabolism	1	1
K01227	E3.2.1.96	mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase [EC:3.2.1.96]	Glycan Biosynthesis and Metabolism	8	2
K12373	HEXA_B	hexosaminidase [EC:3.2.1.52]	Carbohydrate Metabolism - Folding, Sorting and Degradation - Glycan Biosynthesis and Metabolism - Transport and Catabolism	1	1
ko00520_Amino sugar and nucleotide sugar metabolism					
K00012	UGDH, ugd	UDPglucose 6-dehydrogenase [EC:1.1.1.22]	Carbohydrate Metabolism	91	5
K00066	algD	GDP-mannose 6-dehydrogenase [EC:1.1.1.132]	Carbohydrate Metabolism - Signal Transduction	3	1
K00523	ascD, ddhD, rfbI	CDP-4-dehydro-6-deoxyglucose reductase [EC:1.17.1.1]	Carbohydrate Metabolism	7	2
K00790	murA	UDP-N-acetylglucosamine 1-carboxyvinyltransferase [EC:2.5.1.7]	Carbohydrate Metabolism - Glycan Biosynthesis and Metabolism	12	2
K00820	E2.6.1.16, glmS	glucosamine--fructose-6-phosphate aminotransferase (isomerizing) [EC:2.6.1.16]	Carbohydrate Metabolism - Enzyme Families - Amino Acid Metabolism	20	4
K00845	glk	glucokinase [EC:2.7.1.2]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	2	2
K00847	E2.7.1.4, scrK	fructokinase [EC:2.7.1.4]	Carbohydrate Metabolism	1	1
K00966	GMPP	mannose-1-phosphate guanylyltransferase [EC:2.7.7.13]	Carbohydrate Metabolism	4	2
K00971	E2.7.7.22, manC	mannose-1-phosphate guanylyltransferase [EC:2.7.7.22]	Carbohydrate Metabolism	12	4
K00978	rfbF	glucose-1-phosphate cytidylyltransferase [EC:2.7.7.33]	Carbohydrate Metabolism	6	1

K00983	E2.7.7.43, neuA, CMAS	N-acylneuraminate cytidylyltransferase [EC:2.7.7.43]	Carbohydrate Metabolism	11	5
K01183	E3.2.1.14	chitinase [EC:3.2.1.14]	Carbohydrate Metabolism	1	1
K01207	E3.2.1.52, nagZ	beta-N-acetylhexosaminidase [EC:3.2.1.52]	Carbohydrate Metabolism	1	1
K01209	E3.2.1.55, abfA	alpha-N-arabinofuranosidase [EC:3.2.1.55]	Carbohydrate Metabolism	1	1
K01654	E2.5.1.56, neuB	N-acetylneuraminate synthase [EC:2.5.1.56]	Carbohydrate Metabolism	126	8
K01709	rfbG	CDP-glucose 4,6-dehydratase [EC:4.2.1.45]	Carbohydrate Metabolism	6	3
K01711	E4.2.1.47, gmd	GDPmannose 4,6-dehydratase [EC:4.2.1.47]	Carbohydrate Metabolism	533	7
K01784	galE, GALE	UDP-glucose 4-epimerase [EC:5.1.3.2]	Carbohydrate Metabolism	175	8
K01791	wecB	UDP-N-acetylglucosamine 2-epimerase [EC:5.1.3.14]	Carbohydrate Metabolism - Glycan Biosynthesis and Metabolism	38	7
K01809	E5.3.1.8, manA	mannose-6-phosphate isomerase [EC:5.3.1.8]	Carbohydrate Metabolism	17	3
K01810	GPI, pgi	glucose-6-phosphate isomerase [EC:5.3.1.9]	Carbohydrate Metabolism	2	2
K01840	E5.4.2.8, manB	phosphomannomutase [EC:5.4.2.8]	Carbohydrate Metabolism	1	1
K02377	E1.1.1.271, fcl	GDP-L-fucose synthase [EC:1.1.1.271]	Carbohydrate Metabolism	251	7
K02472	wecC	UDP-N-acetyl-D-mannosaminuronic acid dehydrogenase [EC:1.1.1.-]	Carbohydrate Metabolism	1	1
K02473	E5.1.3.7, wbpP	UDP-N-acetylglucosamine 4-epimerase [EC:5.1.3.7]	Carbohydrate Metabolism	2	2
K02564	nagB, GNPDA	glucosamine-6-phosphate deaminase [EC:3.5.99.6]	Carbohydrate Metabolism	2	2
K02795	PTS-Man-EIIC, manY	PTS system, mannose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
K03431	glmM	phosphoglucosamine mutase [EC:5.4.2.10]	Carbohydrate Metabolism	5	2
K04042	glmU	bifunctional UDP-N-acetylglucosamine pyrophosphorylase / Glucosamine-1-phosphate N-acetyltransferase [EC:2.7.7.23 2.3.1.157]	Carbohydrate Metabolism	2	2
K05304	NANS, SAS	sialic acid synthase [EC:2.5.1.56 2.5.1.57]	Carbohydrate Metabolism	1	1
K06118	E3.13.1.1, sqd1, sqdb	UDP-sulfoquinovose synthase [EC:3.13.1.1]	Carbohydrate Metabolism - Lipid Metabolism	2	2
K07106	murQ	N-acetylmuramic acid 6-phosphate etherase [EC:4.2.-.-]	Carbohydrate Metabolism	1	1
K07806	arnB, pmrH	UDP-4-amino-4-deoxy-L-arabinose-oxoglutarate aminotransferase [EC:2.6.1.87]	Carbohydrate Metabolism - Signal Transduction - Glycan Biosynthesis and Metabolism - Amino Acid Metabolism	3	2
K11528	glmU1	UDP-N-acetylglucosamine pyrophosphorylase [EC:2.7.7.23]	Carbohydrate Metabolism	1	1
K12373	HEXA_B	hexosaminidase [EC:3.2.1.52]	Carbohydrate Metabolism - Folding, Sorting and Degradation - Glycan Biosynthesis and Metabolism - Transport and Catabolism	1	1
K12410	npdA	NAD-dependent deacetylase [EC:3.5.1.-]	Carbohydrate Metabolism	4	1
K12452	rfbH	CDP-6-deoxy-D-xylo-4-hexulose-3-dehydrase	Carbohydrate Metabolism	329	6
K12454	rfbE	CDP-paratose 2-epimerase [EC:5.1.3.10]	Carbohydrate Metabolism	19	4

ko00521_Streptomycin biosynthesis

K00067	rfbD	dTDP-4-dehydrorhamnose reductase [EC:1.1.1.133]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	39	4
K00845	glk	glucokinase [EC:2.7.1.2]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	2	2
K00973	E2.7.7.24, rfbA, rffH	glucose-1-phosphate thymidylyltransferase [EC:2.7.7.24]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	28	5
K01092	E3.1.3.25, IMPA, suhB	myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]	Carbohydrate Metabolism - Signal Transduction - Biosynthesis of Other Secondary Metabolites	3	2
K01710	E4.2.1.46, rfbB, rffG	dTDP-glucose 4,6-dehydratase [EC:4.2.1.46]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	341	7
K01790	rfbC	dTDP-4-dehydrorhamnose 3,5-epimerase [EC:5.1.3.13]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	89	4
K01858	E5.5.1.4, INO1	myo-inositol-1-phosphate synthase [EC:5.5.1.4]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	1	1
K04340	strB1	scyllo-inosamine-4-phosphate amidinotransferase 1 [EC:2.1.4.2]	Biosynthesis of Other Secondary Metabolites	16	2

ko00523_Polyketide sugar unit biosynthesis

K00067	rfbD	dTDP-4-dehydrorhamnose reductase [EC:1.1.1.133]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	39	4
K00973	E2.7.7.24, rfbA, rffH	glucose-1-phosphate thymidylyltransferase [EC:2.7.7.24]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	28	5
K01710	E4.2.1.46, rfbB, rffG	dTDP-glucose 4,6-dehydratase [EC:4.2.1.46]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	341	7
K01790	rfbC	dTDP-4-dehydrorhamnose 3,5-epimerase [EC:5.1.3.13]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	89	4

ko00524_Butirosin and neomycin biosynthesis

K00845	glk	glucokinase [EC:2.7.1.2]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	2	2

ko00531_Glycosaminoglycan degradation

K12373	HEXA_B	hexosaminidase [EC:3.2.1.52]	Carbohydrate Metabolism - Folding, Sorting and Degradation - Glycan Biosynthesis and Metabolism - Transport and Catabolism	1	1

ko00540_Lipopolysaccharide biosynthesis

K00677	lpxA	UDP-N-acetylglucosamine acyltransferase [EC:2.3.1.129]	Glycan Biosynthesis and Metabolism	137	3
K00979	kdsB	3-deoxy-manno-octulosonate cytidylyltransferase (CMP-KDO synthetase) [EC:2.7.7.38]	Glycan Biosynthesis and Metabolism	4	2
K01627	kdsA	2-dehydro-3-deoxyphosphooctonate aldolase (KDO 8-P synthase) [EC:2.5.1.55]	Glycan Biosynthesis and Metabolism	24	5
K02535	lpxC	UDP-3-O-[3-hydroxymyristoyl] N-acetylglucosamine deacetylase [EC:3.5.1.-]	Glycan Biosynthesis and Metabolism	5	2
K02536	lpxD	UDP-3-O-[3-hydroxymyristoyl] glucosamine N-acyltransferase [EC:2.3.1.-]	Glycan Biosynthesis and Metabolism	12	2
K02841	waaC, rfaC	heptosyltransferase I [EC:2.4.-.-]	Glycan Biosynthesis and Metabolism	3	2
K02843	waaF, rfaF	heptosyltransferase II [EC:2.4.-.-]	Glycan Biosynthesis and Metabolism	18	2
K02849	waaQ, rfaQ	heptosyltransferase III [EC:2.4.-.-]	Glycan Biosynthesis and Metabolism	1	1
K03270	kdsC	3-deoxy-D-manno-octulosonate 8-phosphate phosphatase (KDO 8-P phosphatase) [EC:3.1.3.45]	Glycan Biosynthesis and Metabolism	6	1
K03271	gmhA	phosphoheptose isomerase [EC:5.-.-.-]	Glycan Biosynthesis and Metabolism	4	3
K03272	gmhC, hldE, waaE, rfaE	D-beta-D-heptose 7-phosphate kinase / D-beta-D-heptose 1-phosphate adenosyltransferase [EC:2.7.1.- 2.7.7.-]	Glycan Biosynthesis and Metabolism	71	4
K03273	gmhB	D-glycero-D-manno-heptose 1,7-bisphosphate phosphatase [EC:3.1.3.-]	Glycan Biosynthesis and Metabolism	1	1
K03274	rfaD	ADP-L-glycero-D-manno-heptose 6-epimerase [EC:5.1.3.20]	Glycan Biosynthesis and Metabolism	46	3

ko00550_Peptidoglycan biosynthesis

K00790	murA	UDP-N-acetylglucosamine 1-carboxyvinyltransferase [EC:2.5.1.7]	Carbohydrate Metabolism - Glycan Biosynthesis and Metabolism	12	2
K01000	mraY	phospho-N-acetylmuramoyl-pentapeptide-transferase [EC:2.7.8.13]	Glycan Biosynthesis and Metabolism	2	1
K01921	ddl	D-alanine-D-alanine ligase [EC:6.3.2.4]	Metabolism of Other Amino Acids - Glycan Biosynthesis and Metabolism	2	2
K01925	murD	UDP-N-acetylmuramoylalanine--D-glutamate ligase [EC:6.3.2.9]	Metabolism of Other Amino Acids - Glycan Biosynthesis and Metabolism	2	2
K01929	murF	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate--D-alanyl-D-alanine ligase [EC:6.3.2.10]	Glycan Biosynthesis and Metabolism - Amino Acid Metabolism	1	1
K03587	ftsI	cell division protein FtsI (penicillin-binding protein 3) [EC:2.4.1.129]	Glycan Biosynthesis and Metabolism - Replication and Repair	1	1
K05366	mrcA	penicillin-binding protein 1A [EC:2.4.1.- 3.4.-.-]	Glycan Biosynthesis and Metabolism	1	1
K05367	pbpC	penicillin-binding protein 1C [EC:2.4.1.-]	Glycan Biosynthesis and Metabolism	2	1
K06153	E3.6.1.27, bacA	undecaprenyl-diphosphatase [EC:3.6.1.27]	Glycan Biosynthesis and Metabolism	1	1
K07260	vanY	D-alanyl-D-alanine carboxypeptidase [EC:3.4.16.4]	Enzyme Families - Glycan Biosynthesis and Metabolism	6	2

ko00561_Glycerolipid metabolism

K00005	E1.1.1.6, gldA	glycerol dehydrogenase [EC:1.1.1.6]	Lipid Metabolism	1	1
K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00655	plsC	1-acyl-sn-glycerol-3-phosphate acyltransferase [EC:2.3.1.51]	Lipid Metabolism	8	2
K00864	E2.7.1.30, glpK	glycerol kinase [EC:2.7.1.30]	Environmental Adaptation - Endocrine System - Lipid Metabolism	3	3
K01046	E3.1.1.3	triacylglycerol lipase [EC:3.1.1.3]	Lipid Metabolism	7	3
K03429	ugtP	1,2-diacylglycerol 3-glucosyltransferase [EC:2.4.1.157]	Glycan Biosynthesis and Metabolism - Lipid Metabolism	8	1
K06118	E3.13.1.1, sqd1, sqdb	UDP-sulfoquinovose synthase [EC:3.13.1.1]	Carbohydrate Metabolism - Lipid Metabolism	2	2

ko00562_Inositol phosphate metabolism

K00140	E1.2.1.27, mmsA, iolA	methylmalonate-semialdehyde dehydrogenase [EC:1.2.1.27]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K01092	E3.1.3.25, IMPA, suhB	myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]	Carbohydrate Metabolism - Signal Transduction - Biosynthesis of Other Secondary Metabolites	3	2
K01858	E5.5.1.4, INO1	myo-inositol-1-phosphate synthase [EC:5.5.1.4]	Carbohydrate Metabolism - Biosynthesis of Other Secondary Metabolites	1	1
K03338	iolC	5-dehydro-2-deoxygluconokinase [EC:2.7.1.92]	Carbohydrate Metabolism	1	1

ko00564_Glycerophospholipid metabolism

K00057	gpsA	glycerol-3-phosphate dehydrogenase (NAD(P)+) [EC:1.1.1.94]	Lipid Metabolism	2	2

K00111	glpA, glpD	glycerol-3-phosphate dehydrogenase [EC:1.1.5.3]	Lipid Metabolism	1	1
K00655	plsC	1-acyl-sn-glycerol-3-phosphate acyltransferase [EC:2.3.1.51]	Lipid Metabolism	8	2
K00968	E2.7.7.15, PCYT1	choline-phosphate cytidylyltransferase [EC:2.7.7.15]	Metabolism of Other Amino Acids - Lipid Metabolism	12	3
K00980	tagD	glycerol-3-phosphate cytidylyltransferase [EC:2.7.7.39]	Lipid Metabolism	86	6
K00995	E2.7.8.5, pgsA	CDP-diacylglycerol--glycerol-3-phosphate 3-phosphatidyltransferase [EC:2.7.8.5]	Lipid Metabolism	1	1
K00998	E2.7.8.8, pssA	phosphatidylserine synthase [EC:2.7.8.8]	Amino Acid Metabolism - Lipid Metabolism	2	2
K01096	pgpB	phosphatidylglycerophosphatase B [EC:3.1.3.27]	Lipid Metabolism	1	1
K01613	E4.1.1.65, psd	phosphatidylserine decarboxylase [EC:4.1.1.65]	Lipid Metabolism	27	4

ko00590_Arachidonic acid metabolism

K00432	E1.11.1.9	glutathione peroxidase [EC:1.11.1.9]	Metabolism of Other Amino Acids - Lipid Metabolism	2	2

ko00591_Linoleic acid metabolism

K00100	E1.1.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Lipid Metabolism	1	1

ko00600_Sphingolipid metabolism

K01190	lacZ	beta-galactosidase [EC:3.2.1.23]	Carbohydrate Metabolism - Glycan Biosynthesis and Metabolism - Lipid Metabolism	1	1

ko00603_Glycosphingolipid biosynthesis - globo series

K12373	HEXA_B	hexosaminidase [EC:3.2.1.52]	Carbohydrate Metabolism - Folding, Sorting and Degradation - Glycan Biosynthesis and Metabolism - Transport and Catabolism	1	1

ko00604_Glycosphingolipid biosynthesis - ganglio series

K12373	HEXA_B	hexosaminidase [EC:3.2.1.52]	Carbohydrate Metabolism - Folding, Sorting and Degradation - Glycan Biosynthesis and Metabolism - Transport and Catabolism	1	1

ko00620_Pyruvate metabolism

K00024	mdh	malate dehydrogenase [EC:1.1.1.37]	Carbohydrate Metabolism - Energy Metabolism	3	3
K00027	E1.1.1.38, sfcA, maeA	malate dehydrogenase (oxaloacetate-decarboxylating) [EC:1.1.1.38]	Carbohydrate Metabolism - Signal Transduction	2	1
K00029	E1.1.1.40, maeB	malate dehydrogenase (oxaloacetate-decarboxylating)(NADP+) [EC:1.1.1.40]	Carbohydrate Metabolism - Energy Metabolism	2	1
K00101	E1.1.2.3, lldD	L-lactate dehydrogenase (cytochrome) [EC:1.1.2.3]	Carbohydrate Metabolism	1	1
K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00156	poxB	pyruvate dehydrogenase (quinone) [EC:1.2.5.1]	Carbohydrate Metabolism	3	2
K00161	PDHA, pdhA	pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	57	3
K00162	PDHB, pdhB	pyruvate dehydrogenase E1 component subunit beta [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	37	3
K00382	DLD, lpd, pdhD	dihydrolipoamide dehydrogenase [EC:1.8.1.4]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00627	DLAT, aceF, pdhC	pyruvate dehydrogenase E2 component (dihydrolipoamide acetyltransferase) [EC:2.3.1.12]	Carbohydrate Metabolism	1	1
K01006	ppdK	pyruvate,orthophosphate dikinase [EC:2.7.9.1]	Carbohydrate Metabolism - Energy Metabolism	1	1
K01007	pps, ppsA	pyruvate, water dikinase [EC:2.7.9.2]	Carbohydrate Metabolism - Energy Metabolism	107	4
K01595	ppc	phosphoenolpyruvate carboxylase [EC:4.1.1.31]	Carbohydrate Metabolism - Energy Metabolism	1	1
K01610	E4.1.1.49, pckA	phosphoenolpyruvate carboxykinase (ATP) [EC:4.1.1.49]	Carbohydrate Metabolism - Energy Metabolism	4	2
K01638	E2.3.3.9, aceB, glcB	malate synthase [EC:2.3.3.9]	Carbohydrate Metabolism	9	3
K01649	E2.3.3.13, leuA	2-isopropylmalate synthase [EC:2.3.3.13]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K01734	E4.2.3.3, mgsA	methylglyoxal synthase [EC:4.2.3.3]	Carbohydrate Metabolism	30	3
K01759	E4.4.1.5, GLO1, gloA	lactoylglutathione lyase [EC:4.4.1.5]	Carbohydrate Metabolism - Signal Transduction	8	1
K01895	ACSS, acs	acetyl-CoA synthetase [EC:6.2.1.1]	Carbohydrate Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01962	accA	acetyl-CoA carboxylase carboxyl transferase subunit alpha [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	1	1
K01963	accD	acetyl-CoA carboxylase carboxyl transferase subunit beta [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	2	2
K03778	ldhA	D-lactate dehydrogenase [EC:1.1.1.28]	Carbohydrate Metabolism	1	1
K04073	mhpF	acetaldehyde dehydrogenase [EC:1.2.1.10]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism	43	2

ko00621_Dioxin degradation

K01666	E4.1.3.39, mhpE	4-hydroxy 2-oxovalerate aldolase [EC:4.1.3.39]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	35	3
K02554	mhpD	2-keto-4-pentenoate hydratase [EC:4.2.1.80]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	1	1
K04073	mhpF	acetaldehyde dehydrogenase [EC:1.2.1.10]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism	43	2

ko00622_Xylene degradation

K01666	E4.1.3.39, mhpE	4-hydroxy 2-oxovalerate aldolase [EC:4.1.3.39]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	35	3
K02554	mhpD	2-keto-4-pentenoate hydratase [EC:4.2.1.80]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	1	1
K04073	mhpF	acetaldehyde dehydrogenase [EC:1.2.1.10]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism	43	2

ko00623_Toluene degradation

K00239	sdhA	succinate dehydrogenase flavoprotein subunit [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	2	2
K00240	sdhB	succinate dehydrogenase iron-sulfur protein [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	18	1

ko00624_Polycyclic aromatic hydrocarbon degradation

K00599	E2.1.1.-		Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	55	3

ko00625_Chloroalkane and chloroalkene degradation

K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00100	E1.1.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Lipid Metabolism	1	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K01563	dhaA	haloalkane dehalogenase [EC:3.8.1.5]	Xenobiotics Biodegradation and Metabolism	1	1
K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1

ko00626_Naphthalene degradation

K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K00257	E1.3.99.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides	2	2
K00680	E2.3.1.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism	9	1
K01726	E4.2.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides	1	1
K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1

ko00627_Aminobenzoate degradation

K00680	E2.3.1.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism	9	1
K01426	E3.5.1.4, amiE	amidase [EC:3.5.1.4]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	3	1
K01501	E3.5.5.1	nitrilase [EC:3.5.5.1]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1

K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K09474	phoN	acid phosphatase (class A) [EC:3.1.3.2]	Signal Transduction - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism	8	2
ko00630_Glyoxylate and dicarboxylate metabolism					
K00023	E1.1.1.36, phbB	acetoacetyl-CoA reductase [EC:1.1.1.36]	Carbohydrate Metabolism	3	2
K00024	mdh	malate dehydrogenase [EC:1.1.1.37]	Carbohydrate Metabolism - Energy Metabolism	3	3
K00050	E1.1.1.81, ttuD	hydroxypyruvate reductase [EC:1.1.1.81]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K00104	glcD	glycolate oxidase [EC:1.1.3.15]	Carbohydrate Metabolism	1	1
K00123	E1.2.1.2A	formate dehydrogenase, alpha subunit [EC:1.2.1.2]	Carbohydrate Metabolism - Energy Metabolism	2	2
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K01091	E3.1.3.18, gph	phosphoglycolate phosphatase [EC:3.1.3.18]	Carbohydrate Metabolism	1	1
K01433	purU	formyltetrahydrofolate deformylase [EC:3.5.1.10]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins	1	1
K01601	rbcL	ribulose-bisphosphate carboxylase large chain [EC:4.1.1.39]	Carbohydrate Metabolism - Energy Metabolism	6	3
K01637	E4.1.3.1, aceA	isocitrate lyase [EC:4.1.3.1]	Carbohydrate Metabolism	9	3
K01638	E2.3.3.9, aceB, glcB	malate synthase [EC:2.3.3.9]	Carbohydrate Metabolism	9	3
K01681	ACO, acnA	aconitate hydratase 1 [EC:4.2.1.3]	Carbohydrate Metabolism - Energy Metabolism	4	3
K01847	MUT	methylmalonyl-CoA mutase [EC:5.4.99.2]	Carbohydrate Metabolism - Amino Acid Metabolism - Energy Metabolism	2	2
K11472	glcE	glycolate oxidase FAD binding subunit	Carbohydrate Metabolism	2	1
ko00633_Nitrotoluene degradation					
K06281	E1.12.99.6L	hydrogenase large subunit [EC:1.12.99.6]	Xenobiotics Biodegradation and Metabolism	1	1
K11180	dsrA	sulfite reductase, dissimilatory-type alpha subunit [EC:1.8.99.3]	Xenobiotics Biodegradation and Metabolism	2	1
ko00640_Propanoate metabolism					
K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00140	E1.2.1.27, mmsA, iolA	methylmalonate-semialdehyde dehydrogenase [EC:1.2.1.27]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
K00249	E1.3.99.3, ACADM, acd	acyl-CoA dehydrogenase [EC:1.3.99.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Endocrine System - Amino Acid Metabolism - Lipid Metabolism	1	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00823	puuE	4-aminobutyrate aminotransferase [EC:2.6.1.19]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K01604	mmdA	methylmalonyl-CoA decarboxylase alpha chain [EC:4.1.1.41]	Carbohydrate Metabolism	1	1
K01659	E2.3.3.5, prpC	2-methylcitrate synthase [EC:2.3.3.5]	Carbohydrate Metabolism	1	1
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01847	MUT	methylmalonyl-CoA mutase [EC:5.4.99.2]	Carbohydrate Metabolism - Amino Acid Metabolism - Energy Metabolism	2	2
K01848	E5.4.99.2A, mcmA1	methylmalonyl-CoA mutase, N-terminal domain [EC:5.4.99.2]	Carbohydrate Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K01895	ACSS, acs	acetyl-CoA synthetase [EC:6.2.1.1]	Carbohydrate Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01962	accA	acetyl-CoA carboxylase carboxyl transferase subunit alpha [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	1	1
K01963	accD	acetyl-CoA carboxylase carboxyl transferase subunit beta [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	2	2
K07250	gabT	4-aminobutyrate aminotransferase / (S)-3-amino-2-methylpropionate transaminase [EC:2.6.1.19 2.6.1.22]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
ko00642_Ethylbenzene degradation					
K00529	hcaD	ferredoxin--NAD+ reductase [EC:1.18.1.3]	Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00680	E2.3.1.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism	9	1
ko00643_Styrene degradation					
K01426	E3.5.1.4, amiE	amidase [EC:3.5.1.4]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	3	1
K01501	E3.5.5.1	nitrilase [EC:3.5.5.1]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1
ko00650_Butanoate metabolism					
K00023	E1.1.1.36, phbB	acetoacetyl-CoA reductase [EC:1.1.1.36]	Carbohydrate Metabolism	3	2
K00100	E1.1.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Lipid Metabolism	1	1
K00135	E1.2.1.16, gabD	succinate-semialdehyde dehydrogenase (NADP+) [EC:1.2.1.16]	Carbohydrate Metabolism - Amino Acid Metabolism	2	1
K00161	PDHA, pdhA	pyruvate dehydrogenase E1 component subunit alpha [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	57	3
K00162	PDHB, pdhB	pyruvate dehydrogenase E1 component subunit beta [EC:1.2.4.1]	Carbohydrate Metabolism - Amino Acid Metabolism	37	3
K00239	sdhA	succinate dehydrogenase flavoprotein subunit [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	2	2
K00240	sdhB	succinate dehydrogenase iron-sulfur protein [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	18	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00823	puuE	4-aminobutyrate aminotransferase [EC:2.6.1.19]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
K01640	E4.1.3.4, HMGCL, hmgL	hydroxymethylglutaryl-CoA lyase [EC:4.1.3.4]	Carbohydrate Metabolism - Transport and Catabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	18	1
K01652	E2.2.1.6L, ilvB, ilvG, ilvI	acetolactate synthase I/II/III large subunit [EC:2.2.1.6]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	300	5
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01726	E4.2.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides	1	1
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K03821	phbC, phaC	polyhydroxyalkanoate synthase [EC:2.3.1.-]	Carbohydrate Metabolism	1	1
K04073	mhpF	acetaldehyde dehydrogenase [EC:1.2.1.10]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism	43	2
K07250	gabT	4-aminobutyrate aminotransferase / (S)-3-amino-2-methylpropionate transaminase [EC:2.6.1.19 2.6.1.22]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Amino Acid Metabolism	1	1
ko00660_C5-Branched dibasic acid metabolism					
K01652	E2.2.1.6L, ilvB, ilvG, ilvI	acetolactate synthase I/II/III large subunit [EC:2.2.1.6]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	300	5
K01703	leuC	3-isopropylmalate/(R)-2-methylmalate dehydratase large subunit [EC:4.2.1.33 4.2.1.35]	Carbohydrate Metabolism - Amino Acid Metabolism	2	2
K01704	leuD	3-isopropylmalate/(R)-2-methylmalate dehydratase small subunit [EC:4.2.1.33 4.2.1.35]	Carbohydrate Metabolism - Amino Acid Metabolism	1	1
ko00670_One carbon pool by folate					
K00287	folA	dihydrofolate reductase [EC:1.5.1.3]	Metabolism of Cofactors and Vitamins	4	3
K00297	E1.5.1.20, metF	methylenetetrahydrofolate reductase (NADPH) [EC:1.5.1.20]	Metabolism of Cofactors and Vitamins - Energy Metabolism	1	1
K00548	E2.1.1.13, metH	5-methyltetrahydrofolate--homocysteine methyltransferase [EC:2.1.1.13]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	23	5
K00560	E2.1.1.45, thyA	thymidylate synthase [EC:2.1.1.45]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	528	9
K00600	E2.1.2.1, glyA	glycine hydroxymethyltransferase [EC:2.1.2.1]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	25	4
K00604	MTFMT, fmt	methionyl-tRNA formyltransferase [EC:2.1.2.9]	Translation - Metabolism of Cofactors and Vitamins	50	5
K00605	E2.1.2.10, gcvT	aminomethyltransferase [EC:2.1.2.10]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	2	2
K01433	purU	formyltetrahydrofolate deformylase [EC:3.5.1.10]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins	1	1
K01938	fhs	formate--tetrahydrofolate ligase [EC:6.3.4.3]	Metabolism of Cofactors and Vitamins - Energy Metabolism	1	1

K03465	E2.1.1.148, thyX, thy1	thymidylate synthase (FAD) [EC:2.1.1.148]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	1150	7
K11175	purN	phosphoribosylglycinamide formyltransferase 1 [EC:2.1.2.2]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	12	2
ko00680_Methane metabolism					
K00024	mdh	malate dehydrogenase [EC:1.1.1.37]	Carbohydrate Metabolism - Energy Metabolism	3	3
K00058	serA, PHGDH	D-3-phosphoglycerate dehydrogenase [EC:1.1.1.95]	Amino Acid Metabolism - Energy Metabolism	2	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K00123	E1.2.1.2A	formate dehydrogenase, alpha subunit [EC:1.2.1.2]	Carbohydrate Metabolism - Energy Metabolism	2	2
K00201	E1.2.99.5B, fwdB, fmdB	formylmethanofuran dehydrogenase subunit B [EC:1.2.99.5]	Energy Metabolism	1	1
K00297	E1.5.1.20, metF	methylenetetrahydrofolate reductase (NADPH) [EC:1.5.1.20]	Metabolism of Cofactors and Vitamins - Energy Metabolism	1	1
K00600	E2.1.2.1, glyA	glycine hydroxymethyltransferase [EC:2.1.2.1]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	25	4
K00831	serC, PSAT1	phosphoserine aminotransferase [EC:2.6.1.52]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	2	1
K01007	pps, ppsA	pyruvate, water dikinase [EC:2.7.9.2]	Carbohydrate Metabolism - Energy Metabolism	107	4
K01079	serB, PSPH	phosphoserine phosphatase [EC:3.1.3.3]	Amino Acid Metabolism - Energy Metabolism	4	3
K01595	ppc	phosphoenolpyruvate carboxylase [EC:4.1.1.31]	Carbohydrate Metabolism - Energy Metabolism	1	1
K01624	FBA, fbaA	fructose-bisphosphate aldolase, class II [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	2	1
K01895	ACSS, acs	acetyl-CoA synthetase [EC:6.2.1.1]	Carbohydrate Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K02118	ATPVB, ntpB	V-type H+-transporting ATPase subunit B [EC:3.6.3.14]	Energy Metabolism	1	1
K02203	thrH	phosphoserine / homoserine phosphotransferase [EC:3.1.3.3 2.7.1.39]	Amino Acid Metabolism - Energy Metabolism	1	1
K03781	katE, CAT	catalase [EC:1.11.1.6]	Neurodegenerative Diseases - Transport and Catabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K03782	katG	catalase/peroxidase [EC:1.11.1.6 1.11.1.7]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	9	1
K14940	cofF	gamma-F420-2:alpha-L-glutamate ligase [EC:6.3.2.32]	Energy Metabolism	1	1
ko00710_Carbon fixation in photosynthetic organisms					
K00024	mdh	malate dehydrogenase [EC:1.1.1.37]	Carbohydrate Metabolism - Energy Metabolism	3	3
K00029	E1.1.1.40, maeB	malate dehydrogenase (oxaloacetate-decarboxylating)(NADP+) [EC:1.1.1.40]	Carbohydrate Metabolism - Energy Metabolism	2	1
K00615	E2.2.1.1, tktA, tktB	transketolase [EC:2.2.1.1]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Energy Metabolism	59	3
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00927	PGK, pgk	phosphoglycerate kinase [EC:2.7.2.3]	Carbohydrate Metabolism - Energy Metabolism	2	1
K01006	ppdK	pyruvate,orthophosphate dikinase [EC:2.7.9.1]	Carbohydrate Metabolism - Energy Metabolism	1	1
K01595	ppc	phosphoenolpyruvate carboxylase [EC:4.1.1.31]	Carbohydrate Metabolism - Energy Metabolism	1	1
K01601	rbcL	ribulose-bisphosphate carboxylase large chain [EC:4.1.1.39]	Carbohydrate Metabolism - Energy Metabolism	6	3
K01610	E4.1.1.49, pckA	phosphoenolpyruvate carboxykinase (ATP) [EC:4.1.1.49]	Carbohydrate Metabolism - Energy Metabolism	4	2
K01623	ALDO, fbaB	fructose-bisphosphate aldolase, class I [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	12	1
K01624	FBA, fbaA	fructose-bisphosphate aldolase, class II [EC:4.1.2.13]	Carbohydrate Metabolism - Energy Metabolism	2	1
K01783	rpe, RPE	ribulose-phosphate 3-epimerase [EC:5.1.3.1]	Carbohydrate Metabolism - Energy Metabolism	1	1
K01808	E5.3.1.6B, rpiB	ribose 5-phosphate isomerase B [EC:5.3.1.6]	Carbohydrate Metabolism - Energy Metabolism	10	3
K04041	fbp3	fructose-1,6-bisphosphatase III [EC:3.1.3.11]	Carbohydrate Metabolism - Energy Metabolism	1	1
ko00720_Carbon fixation pathways in prokaryotes					
K00024	mdh	malate dehydrogenase [EC:1.1.1.37]	Carbohydrate Metabolism - Energy Metabolism	3	3
K00174	korA	2-oxoglutarate ferredoxin oxidoreductase subunit alpha [EC:1.2.7.3]	Carbohydrate Metabolism - Energy Metabolism	1	1
K00175	korB	2-oxoglutarate ferredoxin oxidoreductase subunit beta [EC:1.2.7.3]	Carbohydrate Metabolism - Energy Metabolism	2	2
K00239	sdhA	succinate dehydrogenase flavoprotein subunit [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	2	2
K00240	sdhB	succinate dehydrogenase iron-sulfur protein [EC:1.3.99.1]	Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Energy Metabolism	18	1
K00297	E1.5.1.20, metF	methylenetetrahydrofolate reductase (NADPH) [EC:1.5.1.20]	Metabolism of Cofactors and Vitamins - Energy Metabolism	1	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K01007	pps, ppsA	pyruvate, water dikinase [EC:2.7.9.2]	Carbohydrate Metabolism - Energy Metabolism	107	4
K01595	ppc	phosphoenolpyruvate carboxylase [EC:4.1.1.31]	Carbohydrate Metabolism - Energy Metabolism	1	1
K01681	ACO, acnA	aconitate hydratase 1 [EC:4.2.1.3]	Carbohydrate Metabolism - Energy Metabolism	4	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01847	MUT	methylmalonyl-CoA mutase [EC:5.4.99.2]	Carbohydrate Metabolism - Amino Acid Metabolism - Energy Metabolism	2	2
K01848	E5.4.99.2A, mcmA1	methylmalonyl-CoA mutase, N-terminal domain [EC:5.4.99.2]	Carbohydrate Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K01895	ACSS, acs	acetyl-CoA synthetase [EC:6.2.1.1]	Carbohydrate Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K01938	fhs	formate--tetrahydrofolate ligase [EC:6.3.4.3]	Metabolism of Cofactors and Vitamins - Energy Metabolism	1	1
K01962	accA	acetyl-CoA carboxylase carboxyl transferase subunit alpha [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	1	1
K01963	accD	acetyl-CoA carboxylase carboxyl transferase subunit beta [EC:6.4.1.2]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Lipid Metabolism - Energy Metabolism	2	2
ko00730_Thiamine metabolism					
K00788	thiE	thiamine-phosphate pyrophosphorylase [EC:2.5.1.3]	Metabolism of Cofactors and Vitamins	88	1
K03707	tenA	thiaminase (transcriptional activator TenA) [EC:3.5.99.2]	Transcription - Metabolism of Cofactors and Vitamins	2	1
K04487	iscS, NFS1	cysteine desulfurase [EC:2.8.1.7]	Folding, Sorting and Degradation - Metabolism of Cofactors and Vitamins	4	3
K11717	sufS	cysteine desulfurase / selenocysteine lyase [EC:2.8.1.7 4.4.1.16]	Metabolism of Other Amino Acids - Metabolism of Cofactors and Vitamins	1	1
ko00740_Riboflavin metabolism					
K01497	ribA, RIB1	GTP cyclohydrolase II [EC:3.5.4.25]	Metabolism of Cofactors and Vitamins	1	1
K09474	phoN	acid phosphatase (class A) [EC:3.1.3.2]	Signal Transduction - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism	8	2
K11753	ribF	riboflavin kinase / FMN adenylyltransferase [EC:2.7.1.26 2.7.7.2]	Metabolism of Cofactors and Vitamins	3	2
K14652	ribBA	3,4-dihydroxy 2-butanone 4-phosphate synthase / GTP cyclohydrolase II [EC:4.1.99.12 3.5.4.25]	Metabolism of Cofactors and Vitamins	4	2
K14656	ribL	FAD synthetase	Metabolism of Cofactors and Vitamins	27	3
ko00750_Vitamin B6 metabolism					
K00275	pdxH, PNPO	pyridoxamine 5'-phosphate oxidase [EC:1.4.3.5]	Metabolism of Cofactors and Vitamins	1	1
K00831	serC, PSAT1	phosphoserine aminotransferase [EC:2.6.1.52]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	2	1
ko00760_Nicotinate and nicotinamide metabolism					
K00278	nadB	L-aspartate oxidase [EC:1.4.3.16]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	1	1
K00325	pntB	NAD(P) transhydrogenase subunit beta [EC:1.6.1.2]	Metabolism of Cofactors and Vitamins	1	1
K00763	E2.4.2.11, pncB	nicotinate phosphoribosyltransferase [EC:2.4.2.11]	Metabolism of Cofactors and Vitamins	11	3
K00969	nadD	nicotinate-nucleotide adenylyltransferase [EC:2.7.7.18]	Metabolism of Cofactors and Vitamins	5	2
K01081	E3.1.3.5	5'-nucleotidase [EC:3.1.3.5]	Signaling Molecules and Interaction - Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	3	3
K01916	nadE	NAD+ synthase [EC:6.3.1.5]	Metabolism of Cofactors and Vitamins - Energy Metabolism	80	5
K03462	E2.4.2.12, PBEF1	nicotinamide phosphoribosyltransferase [EC:2.4.2.12]	Metabolism of Cofactors and Vitamins	2	2
K03517	nadA	quinolinate synthase [EC:2.5.1.72]	Metabolism of Cofactors and Vitamins	3	3
K03783	punA	purine-nucleoside phosphorylase [EC:2.4.2.1]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	1	1
K03784	deoD	purine-nucleoside phosphorylase [EC:2.4.2.1]	Metabolism of Cofactors and Vitamins - Nucleotide Metabolism	2	2
K08281	pncA	nicotinamidase/pyrazinamidase [EC:3.5.1.19 3.5.1.-]	Metabolism of Cofactors and Vitamins	1	1
K13522	K13522, nadM	bifunctional NMN adenylyltransferase/nudix hydrolase [EC:2.7.7.1 3.6.1.-]	Metabolism of Cofactors and Vitamins	9	5
ko00770_Pantothenate and CoA biosynthesis					
K00053	ilvC	ketol-acid reductoisomerase [EC:1.1.1.86]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	2	2
K00606	panB	3-methyl-2-oxobutanoate hydroxymethyltransferase [EC:2.1.2.11]	Metabolism of Cofactors and Vitamins	26	4
K00826	E2.6.1.42, ilvE	branched-chain amino acid aminotransferase [EC:2.6.1.42]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	3	2
K00954	E2.7.7.3A, coaD, kdtB	pantetheine-phosphate adenylyltransferase [EC:2.7.7.3]	Metabolism of Cofactors and Vitamins	9	2
K01652	E2.2.1.6L, ilvB, ilvG, ilvI	acetolactate synthase I/II/III large subunit [EC:2.2.1.6]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	300	5
K02201	E2.7.7.3B	pantetheine-phosphate adenylyltransferase [EC:2.7.7.3]	Metabolism of Cofactors and Vitamins	4	2

ko00780_Biotin metabolism					
K00652	bioF	8-amino-7-oxononanoate synthase [EC:2.3.1.47]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	1	1
K01423	E3.4.-.-		Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	7	4
ko00790_Folate biosynthesis					
K00287	folA	dihydrofolate reductase [EC:1.5.1.3]	Metabolism of Cofactors and Vitamins	4	3
K00796	folP	dihydropteroate synthase [EC:2.5.1.15]	Metabolism of Cofactors and Vitamins	1	1
K01495	E3.5.4.16, folE	GTP cyclohydrolase I [EC:3.5.4.16]	Metabolism of Cofactors and Vitamins	495	8
K01737	E4.2.3.12, ptpS	6-pyruvoyl tetrahydrobiopterin synthase [EC:4.2.3.12]	Metabolism of Cofactors and Vitamins	127	5
K03342	pabBC	para-aminobenzoate synthetase / 4-amino-4-deoxychorismate lyase [EC:2.6.1.85 4.1.3.38]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	1	1
K03639	MOCS1, moaA	molybdenum cofactor biosynthesis protein	Folding, Sorting and Degradation - Metabolism of Cofactors and Vitamins	1	1
K04071	E1.1.1.220	6-pyruvoyltetrahydropterin 2'-reductase [EC:1.1.1.220]	Metabolism of Cofactors and Vitamins	4	1
ko00791_Atrazine degradation					
K01428	ureC	urease subunit alpha [EC:3.5.1.5]	Infectious Diseases - Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	2	1
K01429	ureB	urease subunit beta [EC:3.5.1.5]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	1	1
ko00830_Retinol metabolism					
K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
ko00860_Porphyrin and chlorophyll metabolism					
K00228	hemF, CPOX	coproporphyrinogen III oxidase [EC:1.3.3.3]	Metabolism of Cofactors and Vitamins	4	2
K00510	HMOX, hmuO, ho	heme oxygenase [EC:1.14.99.3]	Digestive System - Metabolism of Cofactors and Vitamins	2	1
K00643	E2.3.1.37, ALAS	5-aminolevulinate synthase [EC:2.3.1.37]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	7	4
K00768	E2.4.2.21, cobU, cobT	nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferase [EC:2.4.2.21]	Metabolism of Cofactors and Vitamins	1	1
K00798	E2.5.1.17, cobO, btuR	cob(I)alamin adenosyltransferase [EC:2.5.1.17]	Metabolism of Cofactors and Vitamins	2	1
K01599	hemE, UROD	uroporphyrinogen decarboxylase [EC:4.1.1.37]	Metabolism of Cofactors and Vitamins	2	1
K01749	hemC, HMBS	hydroxymethylbilane synthase [EC:2.5.1.61]	Metabolism of Cofactors and Vitamins	3	1
K01772	hemH, FECH	ferrochelatase [EC:4.99.1.1]	Metabolism of Cofactors and Vitamins	1	1
K01845	hemL	glutamate-1-semialdehyde 2,1-aminomutase [EC:5.4.3.8]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	7	4
K01885	EARS, gltX	glutamyl-tRNA synthetase [EC:6.1.1.17]	Translation - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	1	1
K02217	ftnA, ftn	ferritin [EC:1.16.3.1]	Metabolism of Cofactors and Vitamins	36	2
K02230	cobN	cobaltochelatase CobN [EC:6.6.1.2]	Metabolism of Cofactors and Vitamins	1	1
K02232	E6.3.5.10, cobQ, cbiP	adenosylcobyric acid synthase [EC:6.3.5.10]	Metabolism of Cofactors and Vitamins	1	1
K02301	cyoE	protoheme IX farnesyltransferase [EC:2.5.1.-]	Metabolism of Cofactors and Vitamins - Metabolism of Terpenoids and Polyketides - Energy Metabolism	2	2
K02495	hemN, hemZ	oxygen-independent coproporphyrinogen III oxidase [EC:1.3.99.22]	Metabolism of Cofactors and Vitamins	3	3
K03404	chlD, bchD	magnesium chelatase subunit D [EC:6.6.1.1]	Metabolism of Cofactors and Vitamins	1	1
K03795	cbiX	sirohydrochlorin cobaltochelatase [EC:4.99.1.3]	Metabolism of Cofactors and Vitamins	1	1
K04034	bchE	anaerobic magnesium-protoporphyrin IX monomethyl ester cyclase [EC:4.-.-.-]	Metabolism of Cofactors and Vitamins	1	1
K05371	pcyA	phycocyanobilin:ferredoxin oxidoreductase [EC:1.3.7.5]	Metabolism of Cofactors and Vitamins	4	2
K09882	cobS	cobaltochelatase CobS [EC:6.6.1.2]	Metabolism of Cofactors and Vitamins	39	2
ko00900_Terpenoid backbone biosynthesis					
K00054	E1.1.1.88	hydroxymethylglutaryl-CoA reductase [EC:1.1.1.88]	Metabolism of Terpenoids and Polyketides	1	1
K00099	dxr	1-deoxy-D-xylulose-5-phosphate reductoisomerase [EC:1.1.1.267]	Metabolism of Terpenoids and Polyketides	1	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00795	ispA	farnesyl diphosphate synthase [EC:2.5.1.1 2.5.1.10]	Metabolism of Terpenoids and Polyketides	1	1
K00806	uppS	undecaprenyl diphosphate synthase [EC:2.5.1.31]	Metabolism of Terpenoids and Polyketides	6	1
K00869	E2.7.1.36, MVK, mvaK1	mevalonate kinase [EC:2.7.1.36]	Transport and Catabolism - Metabolism of Terpenoids and Polyketides	1	1
K00938	E2.7.4.2, mvaK2	phosphomevalonate kinase [EC:2.7.4.2]	Metabolism of Terpenoids and Polyketides	1	1
K01662	dxs	1-deoxy-D-xylulose-5-phosphate synthase [EC:2.2.1.7]	Metabolism of Terpenoids and Polyketides	2	2
K01770	ispF	2-C-methyl-D-erythritol 2,4-cyclodiphosphate synthase [EC:4.6.1.12]	Metabolism of Terpenoids and Polyketides	3	1
K02523	ispB	octaprenyl-diphosphate synthase [EC:2.5.1.90]	Metabolism of Terpenoids and Polyketides	1	1
K03526	E1.17.7.1, gcpE, ispG	(E)-4-hydroxy-3-methylbut-2-enyl-diphosphate synthase [EC:1.17.7.1]	Metabolism of Terpenoids and Polyketides	4	3
K03527	E1.17.1.2, lytB, ispH	4-hydroxy-3-methylbut-2-enyl diphosphate reductase [EC:1.17.1.2]	Metabolism of Terpenoids and Polyketides	2	2
ko00903_Limonene and pinene degradation					
K00128	E1.2.1.3	aldehyde dehydrogenase (NAD+) [EC:1.2.1.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	4	4
K00680	E2.3.1.-		Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism	9	1
K01076	E3.1.2.-		Metabolism of Terpenoids and Polyketides - Lipid Metabolism	1	1
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01726	E4.2.1.-		Carbohydrate Metabolism - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides	1	1
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
ko00906_Carotenoid biosynthesis					
K02291	crtB	phytoene synthase [EC:2.5.1.32]	Metabolism of Terpenoids and Polyketides	1	1
K10027	crtI	phytoene dehydrogenase [EC:1.14.99.-]	Metabolism of Terpenoids and Polyketides	4	2
ko00910_Nitrogen metabolism					
K00261	E1.4.1.3	glutamate dehydrogenase (NAD(P)+) [EC:1.4.1.3]	Metabolism of Other Amino Acids - Excretory System - Amino Acid Metabolism - Energy Metabolism	3	2
K00265	gltB	glutamate synthase (NADPH/NADH) large chain [EC:1.4.1.13 1.4.1.14]	Amino Acid Metabolism - Energy Metabolism	3	3
K00266	gltD	glutamate synthase (NADPH/NADH) small chain [EC:1.4.1.13 1.4.1.14]	Amino Acid Metabolism - Energy Metabolism	1	1
K00285	dadA	D-amino-acid dehydrogenase [EC:1.4.99.1]	Amino Acid Metabolism - Energy Metabolism	19	4
K00362	E1.7.1.4L, nirB	nitrite reductase (NAD(P)H) large subunit [EC:1.7.1.4]	Energy Metabolism	1	1
K00368	E1.7.2.1	nitrite reductase (NO-forming) [EC:1.7.2.1]	Energy Metabolism	1	1
K00370	narG	nitrate reductase 1, alpha subunit [EC:1.7.99.4]	Signal Transduction - Energy Metabolism	1	1
K00372	E1.7.99.4C	nitrate reductase catalytic subunit [EC:1.7.99.4]	Energy Metabolism	1	1
K00459	E1.13.12.16	nitronate monooxygenase [EC:1.13.12.16]	Energy Metabolism	1	1
K00605	E2.1.2.10, gcvT	aminomethyltransferase [EC:2.1.2.10]	Metabolism of Cofactors and Vitamins - Amino Acid Metabolism - Energy Metabolism	2	2
K01424	E3.5.1.1, ansA, ansB	L-asparaginase [EC:3.5.1.1]	Metabolism of Other Amino Acids - Amino Acid Metabolism - Energy Metabolism	1	1
K01501	E3.5.5.1	nitrilase [EC:3.5.5.1]	Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K01667	E4.1.99.1, tnaA	tryptophanase [EC:4.1.99.1]	Amino Acid Metabolism - Energy Metabolism	1	1
K01673	cynT, can	carbonic anhydrase [EC:4.2.1.1]	Energy Metabolism	2	2
K01760	metC	cystathionine beta-lyase [EC:4.4.1.8]	Metabolism of Other Amino Acids - Amino Acid Metabolism - Energy Metabolism	2	1
K01915	E6.3.1.2, glnA	glutamine synthetase [EC:6.3.1.2]	Signal Transduction - Nervous System - Amino Acid Metabolism - Energy Metabolism	113	7
K01916	nadE	NAD+ synthase [EC:6.3.1.5]	Metabolism of Cofactors and Vitamins - Energy Metabolism	80	5
K01953	E6.3.5.4, asnB	asparagine synthase (glutamine-hydrolysing) [EC:6.3.5.4]	Enzyme Families - Amino Acid Metabolism - Energy Metabolism	14	2
K02592	nifN	nitrogenase molybdenum-iron protein NifN	Energy Metabolism	1	1
K04748	norQ	nitric-oxide reductase NorQ protein [EC:1.7.99.7]	Energy Metabolism	15	2

ko00920_Sulfur metabolism					
K00640	E2.3.1.30, cysE	serine O-acetyltransferase [EC:2.3.1.30]	Amino Acid Metabolism - Energy Metabolism	2	2
K00641	E2.3.1.31, metX	homoserine O-acetyltransferase [EC:2.3.1.31]	Amino Acid Metabolism - Energy Metabolism	1	1
K00860	cysC	adenylylsulfate kinase [EC:2.7.1.25]	Nucleotide Metabolism - Energy Metabolism	165	4
K00955	cysNC	bifunctional enzyme CysN/CysC [EC:2.7.7.4 2.7.1.25]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Energy Metabolism	2	2
K00957	cysD	sulfate adenylyltransferase subunit 2 [EC:2.7.7.4]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Energy Metabolism	2	2
K01738	cysK	cysteine synthase A [EC:2.5.1.47]	Amino Acid Metabolism - Energy Metabolism	30	5
K01760	metC	cystathionine beta-lyase [EC:4.4.1.8]	Metabolism of Other Amino Acids - Amino Acid Metabolism - Energy Metabolism	2	1
K12339	cysM	cysteine synthase B [EC:2.5.1.47]	Amino Acid Metabolism - Energy Metabolism	5	2
ko00930_Caprolactam degradation					
K01692	E4.2.1.17, paaG	enoyl-CoA hydratase [EC:4.2.1.17]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	33	3
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
ko00940_Phenylpropanoid biosynthesis					
K00588	E2.1.1.104	caffeoyl-CoA O-methyltransferase [EC:2.1.1.104]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	1	1
K03782	katG	catalase/peroxidase [EC:1.11.1.6 1.11.1.7]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	9	1
K05349	bglX	beta-glucosidase [EC:3.2.1.21]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites	1	1
K05350	bglB	beta-glucosidase [EC:3.2.1.21]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites	4	2
ko00941_Flavonoid biosynthesis					
K00091	E1.1.1.219	dihydroflavonol-4-reductase [EC:1.1.1.219]	Biosynthesis of Other Secondary Metabolites	1	1
K00588	E2.1.1.104	caffeoyl-CoA O-methyltransferase [EC:2.1.1.104]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	1	1
ko00945_Stilbenoid, diarylheptanoid and gingerol biosynthesis					
K00588	E2.1.1.104	caffeoyl-CoA O-methyltransferase [EC:2.1.1.104]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	1	1
ko00950_Isoquinoline alkaloid biosynthesis					
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
ko00960_Tropane, piperidine and pyridine alkaloid biosynthesis					
K00808	hss	homospermidine synthase [EC:2.5.1.44]	Biosynthesis of Other Secondary Metabolites	1	1
K00812	E2.6.1.1A, aspB	aspartate aminotransferase [EC:2.6.1.1]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism - Energy Metabolism	46	3
K00817	hisC	histidinol-phosphate aminotransferase [EC:2.6.1.9]	Biosynthesis of Other Secondary Metabolites - Amino Acid Metabolism	6	3
ko00970_Aminoacyl-tRNA biosynthesis					
K00604	MTFMT, fmt	methionyl-tRNA formyltransferase [EC:2.1.2.9]	Translation - Metabolism of Cofactors and Vitamins	50	5
K01866	YARS, tyrS	tyrosyl-tRNA synthetase [EC:6.1.1.1]	Translation - Amino Acid Metabolism	1	1
K01867	WARS, trpS	tryptophanyl-tRNA synthetase [EC:6.1.1.2]	Translation - Amino Acid Metabolism	3	1
K01869	LARS, leuS	leucyl-tRNA synthetase [EC:6.1.1.4]	Translation - Amino Acid Metabolism	2	2
K01870	IARS, ileS	isoleucyl-tRNA synthetase [EC:6.1.1.5]	Translation - Amino Acid Metabolism	3	2
K01872	AARS, alaS	alanyl-tRNA synthetase [EC:6.1.1.7]	Translation - Amino Acid Metabolism	7	4
K01873	VARS, valS	valyl-tRNA synthetase [EC:6.1.1.9]	Translation - Amino Acid Metabolism	3	2
K01874	MARS, metG	methionyl-tRNA synthetase [EC:6.1.1.10]	Translation - Metabolism of Other Amino Acids - Amino Acid Metabolism	6	3
K01875	SARS, serS	seryl-tRNA synthetase [EC:6.1.1.11]	Translation - Amino Acid Metabolism	3	2
K01876	DARS, aspS	aspartyl-tRNA synthetase [EC:6.1.1.12]	Translation - Amino Acid Metabolism	1	1
K01881	PARS, proS	prolyl-tRNA synthetase [EC:6.1.1.15]	Translation - Amino Acid Metabolism	2	1
K01883	CARS, cysS	cysteinyl-tRNA synthetase [EC:6.1.1.16]	Translation - Amino Acid Metabolism	2	2
K01885	EARS, gltX	glutamyl-tRNA synthetase [EC:6.1.1.17]	Translation - Metabolism of Cofactors and Vitamins - Amino Acid Metabolism	1	1
K01886	QARS, glnS	glutaminyl-tRNA synthetase [EC:6.1.1.18]	Translation - Amino Acid Metabolism	3	1
K01887	RARS, argS	arginyl-tRNA synthetase [EC:6.1.1.19]	Translation - Amino Acid Metabolism	2	1
K01890	FARSB, pheT	phenylalanyl-tRNA synthetase beta chain [EC:6.1.1.20]	Translation - Amino Acid Metabolism	2	1
K01892	HARS, hisS	histidyl-tRNA synthetase [EC:6.1.1.21]	Translation - Amino Acid Metabolism	1	1
K01893	NARS, asnS	asparaginyl-tRNA synthetase [EC:6.1.1.22]	Translation - Amino Acid Metabolism	1	1
K02433	gatA	aspartyl-tRNA(Asn)/glutamyl-tRNA (Gln) amidotransferase subunit A [EC:6.3.5.6 6.3.5.7]	Translation	1	1
K04567	KARS, lysS	lysyl-tRNA synthetase, class II [EC:6.1.1.6]	Translation - Amino Acid Metabolism	18	2
K09482	gatD	glutamyl-tRNA(Gln) amidotransferase subunit D [EC:6.3.5.7]	Translation	1	1
ko00980_Metabolism of xenobiotics by cytochrome P450					
K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
ko00982_Drug metabolism - cytochrome P450					
K00001	E1.1.1.1, adh	alcohol dehydrogenase [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
K00121	frmA, ADH5, adhC	S-(hydroxymethyl)glutathione dehydrogenase / alcohol dehydrogenase [EC:1.1.1.284 1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	8	3
K13953	adhP	alcohol dehydrogenase, propanol-preferring [EC:1.1.1.1]	Carbohydrate Metabolism - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism - Lipid Metabolism	1	1
ko00983_Drug metabolism - other enzymes					
K00088	E1.1.1.205, guaB	IMP dehydrogenase [EC:1.1.1.205]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	57	4
K00760	E2.4.2.8, hpt	hypoxanthine phosphoribosyltransferase [EC:2.4.2.8]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	62	4
K00857	E2.7.1.21, tdk	thymidine kinase [EC:2.7.1.21]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	13	3
K01951	E6.3.5.2, guaA	GMP synthase (glutamine-hydrolysing) [EC:6.3.5.2]	Enzyme Families - Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	1	1
K13421	UMPS	uridine monophosphate synthetase [EC:2.4.2.10 4.1.1.23]	Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism	1	1
ko01040_Biosynthesis of unsaturated fatty acids					
K00059	fabG	3-oxoacyl-[acyl-carrier protein] reductase [EC:1.1.1.100]	Lipid Metabolism	10	4
K00507	SCD, desC	stearoyl-CoA desaturase (delta-9 desaturase) [EC:1.14.19.1]	Endocrine System - Lipid Metabolism	13	2
K01076	E3.1.2.-		Metabolism of Terpenoids and Polyketides - Lipid Metabolism	1	1
K01825	fadB	3-hydroxyacyl-CoA dehydrogenase / enoyl-CoA hydratase / 3-hydroxybutyryl-CoA epimerase / enoyl-CoA isomerase [EC:1.1.1.35 4.2.1.17 5.1.2.3 5.3.3.8]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	1	1
K10806	yciA	acyl-CoA thioesterase YciA [EC:3.1.2.-]	Lipid Metabolism	6	2
ko01051_Biosynthesis of ansamycins					
K00615	E2.2.1.1, tktA, tktB	transketolase [EC:2.2.1.1]	Carbohydrate Metabolism - Metabolism of Terpenoids and Polyketides - Energy Metabolism	59	3
ko01053_Biosynthesis of siderophore group nonribosomal peptides					
K04789	mbtE	mycobactin peptide synthetase MbtE	Metabolism of Terpenoids and Polyketides	2	1
K12240	pchF	pyochelin synthetase	Metabolism of Terpenoids and Polyketides	4	1
ko01055_Biosynthesis of vancomycin group antibiotics					
K01710	E4.2.1.46, rfbB, rffG	dTDP-glucose 4,6-dehydratase [EC:4.2.1.46]	Biosynthesis of Other Secondary Metabolites - Metabolism of Terpenoids and Polyketides	341	7
ko02010_ABC transporters					

K01995	livG	branched-chain amino acid transport system ATP-binding protein	Membrane Transport	1	1
K01996	livF	branched-chain amino acid transport system ATP-binding protein	Membrane Transport	3	2
K01997	livH	branched-chain amino acid transport system permease protein	Membrane Transport	1	1
K01998	livM	branched-chain amino acid transport system permease protein	Membrane Transport	5	3
K01999	livK	branched-chain amino acid transport system substrate-binding protein	Membrane Transport	2	2
K02001	proW	glycine betaine/proline transport system permease protein	Membrane Transport	9	3
K02010	afuC, fbpC	iron(III) transport system ATP-binding protein [EC:3.6.3.30]	Membrane Transport	1	1
K02011	afuB, fbpB	iron(III) transport system permease protein	Membrane Transport	2	1
K02013	ABC.FEV.A	iron complex transport system ATP-binding protein [EC:3.6.3.34]	Membrane Transport	1	1
K02032	ABC.PE.A1	peptide/nickel transport system ATP-binding protein	Membrane Transport	3	2
K02033	ABC.PE.P	peptide/nickel transport system permease protein	Membrane Transport	2	2
K02034	ABC.PE.P1	peptide/nickel transport system permease protein	Membrane Transport	3	3
K02035	ABC.PE.S	peptide/nickel transport system substrate-binding protein	Membrane Transport	8	4
K02040	pstS	phosphate transport system substrate-binding protein	Signal Transduction - Infectious Diseases - Membrane Transport	57	4
K02042	phnE	phosphonate transport system permease protein	Membrane Transport	1	1
K02046	cysU	sulfate transport system permease protein	Membrane Transport	1	1
K02049	ABC.SN.A, ssuB, tauB	sulfonate/nitrate/taurine transport system ATP-binding protein	Membrane Transport	4	3
K02050	ABC.SN.P, ssuC, tauC	sulfonate/nitrate/taurine transport system permease protein	Membrane Transport	3	2
K02072	ABC.MET.P, metI	D-methionine transport system permease protein	Membrane Transport	1	1
K02073	ABC.MET.S, metQ	D-methionine transport system substrate-binding protein	Membrane Transport	1	1
K02194	ccmB	heme exporter protein B	Membrane Transport	1	1
K02195	ccmC	heme exporter protein C	Membrane Transport	1	1
K05685	macB	macrolide transport system ATP-binding/permease protein [EC:3.6.3.-]	Membrane Transport	1	1
K05814	ugpA	sn-glycerol 3-phosphate transport system permease protein	Membrane Transport	1	1
K06857	ABC.TG.A	putative tungstate transport system ATP-binding protein	Membrane Transport	1	1
K06861	lptB	lipopolysaccharide export system ATP-binding protein [EC:3.6.3.-]	Membrane Transport	1	1
K09686	ABC-2.AB.P	antibiotic transport system permease protein	Membrane Transport	4	1
K09687	ABC-2.AB.A	antibiotic transport system ATP-binding protein	Membrane Transport	1	1
K09691	ABC-2.LPSE.A	lipopolysaccharide transport system ATP-binding protein	Membrane Transport	4	2
K09695	nodI	lipooligosaccharide transport system ATP-binding protein	Membrane Transport	1	1
K09808	ABC.LPT.P, lolC, lolE	lipoprotein-releasing system permease protein	Membrane Transport	1	1
K10005	gluB	glutamate transport system substrate-binding protein	Membrane Transport	1	1
K10107	ABC-2.CPSE.P1	capsular polysaccharide transport system permease protein	Membrane Transport	1	1
K10112	msmX, msmK	maltose/maltodextrin transport system ATP-binding protein	Membrane Transport	1	1
K11952	cmpC	bicarbonate transport system ATP-binding protein [EC:3.6.3.-]	Membrane Transport	2	1
K11962	urtD	urea transport system ATP-binding protein	Membrane Transport	1	1

ko02020_Two-component system

K00027	E1.1.1.38, sfcA, maeA	malate dehydrogenase (oxaloacetate-decarboxylating) [EC:1.1.1.38]	Carbohydrate Metabolism - Signal Transduction	2	1
K00066	algD	GDP-mannose 6-dehydrogenase [EC:1.1.1.132]	Carbohydrate Metabolism - Signal Transduction	3	1
K00370	narG	nitrate reductase 1, alpha subunit [EC:1.7.99.4]	Signal Transduction - Energy Metabolism	1	1
K00575	cheR	chemotaxis protein methyltransferase CheR [EC:2.1.1.80]	Cell Motility - Signal Transduction	1	1
K00626	E2.3.1.9, atoB	acetyl-CoA C-acetyltransferase [EC:2.3.1.9]	Carbohydrate Metabolism - Signal Transduction - Xenobiotics Biodegradation and Metabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism - Energy Metabolism	4	3
K00990	glnD	[protein-PII] uridylyltransferase [EC:2.7.7.59]	Signal Transduction	1	1
K01467	E3.5.2.6, ampC, penP	beta-lactamase [EC:3.5.2.6]	Signal Transduction - Biosynthesis of Other Secondary Metabolites	1	1
K01546	kdpA	K+-transporting ATPase ATPase A chain [EC:3.6.3.12]	Signal Transduction	1	1
K01915	E6.3.1.2, glnA	glutamine synthetase [EC:6.3.1.2]	Signal Transduction - Nervous System - Amino Acid Metabolism - Energy Metabolism	113	7
K02040	pstS	phosphate transport system substrate-binding protein	Signal Transduction - Infectious Diseases - Membrane Transport	57	4
K02313	dnaA	chromosomal replication initiator protein	Signal Transduction - Cell Growth and Death - Replication and Repair	470	2
K02405	fliA	RNA polymerase sigma factor for flagellar operon FliA	Cell Motility - Transcription - Signal Transduction - Infectious Diseases	7	2
K02489	pleC1	two-component system, cell cycle sensor kinase and response regulator [EC:2.7.13.3]	Signal Transduction - Enzyme Families	2	1
K02650	pilA	type IV pilus assembly protein PilA	Cell Motility - Signal Transduction - Membrane Transport	5	1
K03092	SIG54, rpoN	RNA polymerase sigma-54 factor	Transcription - Signal Transduction - Infectious Diseases	8	1
K03406	mcp	methyl-accepting chemotaxis protein	Cell Motility - Signal Transduction	74	4
K03407	cheA	two-component system, chemotaxis family, sensor kinase CheA [EC:2.7.13.3]	Cell Motility - Signal Transduction - Enzyme Families	2	2
K03413	cheY	two-component system, chemotaxis family, response regulator CheY	Cell Motility - Signal Transduction	2	2
K03563	csrA	carbon storage regulator	Signal Transduction	5	3
K04751	glnB	nitrogen regulatory protein P-II 1	Signal Transduction	1	1
K06596	chpA	chemosensory pili system protein ChpA (sensor histidine kinase/response regulator)	Cell Motility - Signal Transduction - Enzyme Families	1	1
K07636	phoR	two-component system, OmpR family, phosphate regulon sensor histidine kinase PhoR [EC:2.7.13.3]	Signal Transduction - Enzyme Families	1	1
K07639	rstB	two-component system, OmpR family, sensor histidine kinase RstB [EC:2.7.13.3]	Signal Transduction - Enzyme Families	25	1
K07645	qseC	two-component system, OmpR family, sensor histidine kinase QseC [EC:2.7.13.3]	Signal Transduction - Enzyme Families	7	2
K07649	tctE	two-component system, OmpR family, sensor histidine kinase TctE [EC:2.7.13.3]	Signal Transduction - Enzyme Families	1	1
K07665	cusR	two-component system, OmpR family, copper resistance phosphate regulon response regulator CusR	Signal Transduction	1	1
K07673	narX	two-component system, NarL family, nitrate/nitrite sensor histidine kinase NarX [EC:2.7.13.3]	Signal Transduction - Enzyme Families	1	1
K07714	atoC	two-component system, NtrC family, response regulator AtoC	Signal Transduction	2	2
K07774	tctD	two-component system, OmpR family, response regulator TctD	Signal Transduction	1	1
K07782	sdiA	LuxR family transcriptional regulator	Transcription - Signal Transduction	1	1
K07787	cusA	Cu(I)/Ag(I) efflux system membrane protein CusA	Signal Transduction	2	1
K07788	mdtB	RND superfamily, multidrug transport protein MdtB	Signal Transduction	1	1
K07794	tctB	putative tricarboxylic transport membrane protein	Signal Transduction	1	1
K07799	mdtA	putative multidrug efflux transporter MdtA	Signal Transduction	2	2
K07806	arnB, pmrH	UDP-4-amino-4-deoxy-L-arabinose-oxoglutarate aminotransferase [EC:2.6.1.87]	Carbohydrate Metabolism - Signal Transduction - Glycan Biosynthesis and Metabolism - Amino Acid Metabolism	3	2
K08359	ttrC	tetrathionate reductase subunit C	Signal Transduction	1	1
K09474	phoN	acid phosphatase (class A) [EC:3.1.3.2]	Signal Transduction - Metabolism of Cofactors and Vitamins - Xenobiotics Biodegradation and Metabolism	8	2
K10126	dctD	two-component system, NtrC family, C4-dicarboxylate transport response regulator DctD	Signal Transduction	1	1
K10682	saeR	two-component system, OmpR family, response regulator SaeR	Signal Transduction	1	1
K11331	nrsC	cation efflux system protein involved in nickel and cobalt tolerance [EC:3.2.1.17]	Signal Transduction	24	2

ko02030_Bacterial chemotaxis

K00575	cheR	chemotaxis protein methyltransferase CheR [EC:2.1.1.80]	Cell Motility - Signal Transduction	1	1
K02557	motB	chemotaxis protein MotB	Cell Motility	1	1
K03406	mcp	methyl-accepting chemotaxis protein	Cell Motility - Signal Transduction	74	4
K03407	cheA	two-component system, chemotaxis family, sensor kinase CheA [EC:2.7.13.3]	Cell Motility - Signal Transduction - Enzyme Families	2	2
K03413	cheY	two-component system, chemotaxis family, response regulator CheY	Cell Motility - Signal Transduction	2	2
K03414	cheZ	chemotaxis protein CheZ	Cell Motility	1	1

ko02040_Flagellar assembly

K02389	flgD	flagellar basal-body rod modification protein FlgD	Cell Motility	1	1
K02391	flgF	flagellar basal-body rod protein FlgF	Cell Motility	1	1
K02393	flgH	flagellar L-ring protein precursor FlgH	Cell Motility	8	2
K02396	flgK	flagellar hook-associated protein 1 FlgK	Cell Motility	1	1
K02400	fliA	flagellar biosynthesis protein FlhA	Cell Motility - Membrane Transport	1	1
K02407	fliD	flagellar hook-associated protein 2	Cell Motility	2	2
K02413	fliJ	flagellar FliJ protein	Cell Motility	1	1
K02557	motB	chemotaxis protein MotB	Cell Motility	1	1

ko02060_Phosphotransferase system (PTS)					
K02761	PTS-Cel-EIIC, celB	PTS system, cellobiose-specific IIC component	Membrane Transport	2	1
K02770	PTS-Fru-EIIC, fruA	PTS system, fructose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
K02791	PTS-Mal-EIIC, malX	PTS system, maltose and glucose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
K02795	PTS-Man-EIIC, manY	PTS system, mannose-specific IIC component	Carbohydrate Metabolism - Membrane Transport	1	1
ko03010_Ribosome					
K02877	RP-L15e, RPL15	large subunit ribosomal protein L15e	Translation	1	1
K02890	RP-L22, rplV	large subunit ribosomal protein L22	Translation	1	1
K02896	RP-L24e, RPL24	large subunit ribosomal protein L24e	Translation	1	1
K02912	RP-L32e, RPL32	large subunit ribosomal protein L32e	Translation	1	1
K02913	RP-L33, rpmG	large subunit ribosomal protein L33	Translation	1	1
K02935	RP-L7, rplL	large subunit ribosomal protein L7/L12	Translation	2	1
K02936	RP-L7Ae, RPL7A	large subunit ribosomal protein L7Ae	Translation	3	1
K02945	RP-S1, rpsA	small subunit ribosomal protein S1	Translation	17	3
K02946	RP-S10, rpsJ	small subunit ribosomal protein S10	Translation	1	1
K02948	RP-S11, rpsK	small subunit ribosomal protein S11	Translation	1	1
K02965	RP-S19, rpsS	small subunit ribosomal protein S19	Translation	1	1
K02968	RP-S20, rpsT	small subunit ribosomal protein S20	Translation	4	2
K02970	RP-S21, rpsU	small subunit ribosomal protein S21	Translation	18	4
K02986	RP-S4, rpsD	small subunit ribosomal protein S4	Translation	2	2
K02995	RP-S8e, RPS8	small subunit ribosomal protein S8e	Translation	1	1
ko03013_RNA transport					
K00974	cca	tRNA nucleotidyltransferase (CCA-adding enzyme) [EC:2.7.7.72 3.1.3.-3.1.4.-]	Translation	15	6
ko03018_RNA degradation					
K00962	pnp, PNPT1	polyribonucleotide nucleotidyltransferase [EC:2.7.7.8]	Folding, Sorting and Degradation - Nucleotide Metabolism	1	1
K00970	pcnB	poly(A) polymerase [EC:2.7.7.19]	Folding, Sorting and Degradation	7	4
K04043	dnaK	molecular chaperone DnaK	Folding, Sorting and Degradation - Infectious Diseases	109	6
K04077	groEL, HSPD1	chaperonin GroEL	Metabolic Diseases - Folding, Sorting and Degradation - Infectious Diseases	1783	8
K05592	deaD	ATP-dependent RNA helicase DeaD [EC:3.6.4.13]	Folding, Sorting and Degradation - Translation	1	1
K08300	rne	ribonuclease E [EC:3.1.26.12]	Folding, Sorting and Degradation - Translation	6	2
K12573	rnr, vacB	ribonuclease R [EC:3.1.-.-]	Folding, Sorting and Degradation	3	3
ko03020_RNA polymerase					
K02405	fliA	RNA polymerase sigma factor for flagellar operon FliA	Cell Motility - Transcription - Signal Transduction - Infectious Diseases	7	2
K03040	rpoA	DNA-directed RNA polymerase subunit alpha [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	40	7
K03041	rpoA1	DNA-directed RNA polymerase subunit A' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	5	2
K03042	rpoA2	DNA-directed RNA polymerase subunit A" [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	6	1
K03043	rpoB	DNA-directed RNA polymerase subunit beta [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	5	5
K03044	rpoB1	DNA-directed RNA polymerase subunit B' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	4	3
K03045	rpoB2	DNA-directed RNA polymerase subunit B" [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	2	1
K03046	rpoC	DNA-directed RNA polymerase subunit beta' [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	7	4
K03053	rpoH	DNA-directed RNA polymerase subunit H [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	1	1
K03056	rpoL	DNA-directed RNA polymerase subunit L [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	2	1
K03060	rpoZ	DNA-directed RNA polymerase subunit omega [EC:2.7.7.6]	Transcription - Nucleotide Metabolism - Replication and Repair	1	1
K03086	SIG1, rpoD	RNA polymerase primary sigma factor	Transcription	265	6
K03087	SIG2, rpoS	RNA polymerase nonessential primary-like sigma factor	Transcription - Infectious Diseases	30	3
K03088	SIG3.2, rpoE	RNA polymerase sigma-70 factor, ECF subfamily	Transcription	20	3
K03089	SIG3.3.1, rpoH	RNA polymerase sigma-32 factor	Transcription	2	1
K03090	SIG3.3.2, sigB	RNA polymerase sigma-B factor	Transcription	2	2
K03091	SIG3.4	RNA polymerase sporulation-specific sigma factor	Transcription	2	2
K03092	SIG54, rpoN	RNA polymerase sigma-54 factor	Transcription - Signal Transduction - Infectious Diseases	8	1
K03093	SIGMA70	RNA polymerase sigma factor	Transcription	1	1
K13798	K13798, rpoB	DNA-directed RNA polymerase subunit B [EC:2.7.7.6]	Transcription - Nucleotide Metabolism	3	2
ko03022_Basal transcription factors					
K03120	TBP, tbp	transcription initiation factor TFIID TATA-box-binding protein	Transcription - Neurodegenerative Diseases	3	1
K03124	TFIIB, GTF2B, SUA7, tfb	transcription initiation factor TFIIB	Transcription	11	3
K10843	ERCC3, XPB	DNA excision repair protein ERCC-3 [EC:3.6.4.12]	Transcription - Replication and Repair	4	2
ko03030_DNA replication					
K01972	E6.5.1.2, ligA, ligB	DNA ligase (NAD+) [EC:6.5.1.2]	Replication and Repair	10	2
K02314	dnaB	replicative DNA helicase [EC:3.6.4.12]	Cell Growth and Death - Replication and Repair	1499	9
K02316	dnaG	DNA primase [EC:2.7.7.-]	Replication and Repair	169	6
K02335	DPO1, polA	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	921	8
K02337	DPO3A1, dnaE	DNA polymerase III subunit alpha [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	178	7
K02338	DPO3B, dnaN	DNA polymerase III subunit beta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	18	2
K02340	DPO3D1, holA	DNA polymerase III subunit delta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02341	DPO3D2, holB	DNA polymerase III subunit delta' [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02342	DPO3E, dnaQ	DNA polymerase III subunit epsilon [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	16	3
K02343	DPO3G, dnaX	DNA polymerase III subunit gamma/tau [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	13	3
K03111	ssb	single-strand DNA-binding protein	Replication and Repair	222	8
K03469	E3.1.26.4A, RNASEH1, rnhA	ribonuclease HI [EC:3.1.26.4]	Replication and Repair	3	2
K03470	rnhB	ribonuclease HII [EC:3.1.26.4]	Replication and Repair	3	3
K03763	DPO3A2, polC	DNA polymerase III subunit alpha, Gram-positive type [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	9	3
K10747	LIG1	DNA ligase 1 [EC:6.5.1.1]	Replication and Repair	1	1
ko03050_Proteasome					
K03420	psmR	proteasome regulatory subunit	Folding, Sorting and Degradation	1	1
K03433	psmB, prcB	proteasome beta subunit [EC:3.4.25.1]	Folding, Sorting and Degradation - Enzyme Families	7	2
ko03060_Protein export					
K03070	secA	preprotein translocase subunit SecA	Folding, Sorting and Degradation - Membrane Transport	3	3
K03073	secE	preprotein translocase subunit SecE	Folding, Sorting and Degradation - Membrane Transport	1	1
K03106	SRP54, ffh	signal recognition particle subunit SRP54	Folding, Sorting and Degradation - Membrane Transport	1	1
K03110	ftsY	fused signal recognition particle receptor	Folding, Sorting and Degradation - Membrane Transport	2	2
K03217	yidC, spoIIIJ, OXA1	preprotein translocase subunit YidC	Folding, Sorting and Degradation - Membrane Transport	1	1
ko03070_Bacterial secretion system					
K02453	gspD	general secretion pathway protein D	Membrane Transport	3	2
K03070	secA	preprotein translocase subunit SecA	Folding, Sorting and Degradation - Membrane Transport	3	3
K03073	secE	preprotein translocase subunit SecE	Folding, Sorting and Degradation - Membrane Transport	1	1
K03106	SRP54, ffh	signal recognition particle subunit SRP54	Folding, Sorting and Degradation - Membrane Transport	1	1
K03110	ftsY	fused signal recognition particle receptor	Folding, Sorting and Degradation - Membrane Transport	2	2
K03196	virB11	type IV secretion system protein VirB11	Membrane Transport	2	1
K03217	yidC, spoIIIJ, OXA1	preprotein translocase subunit YidC	Folding, Sorting and Degradation - Membrane Transport	1	1
K03223	yscL	type III secretion protein SctL	Membrane Transport	1	1
K11891	impL, vasK, icmF	type VI secretion system protein ImpL	Membrane Transport	2	2
K11904	vgrG	type VI secretion system secreted protein VgrG	Membrane Transport	5	2
K11907	vasG, clpV	type VI secretion system protein VasG	Membrane Transport	4	2
ko03320_PPAR signaling pathway					
K00249	E1.3.99.3, ACADM, acd	acyl-CoA dehydrogenase [EC:1.3.99.3]	Carbohydrate Metabolism - Metabolism of Other Amino Acids - Endocrine System - Amino Acid Metabolism - Lipid Metabolism	1	1
K00507	SCD, desC	stearoyl-CoA desaturase (delta-9 desaturase) [EC:1.14.19.1]	Endocrine System - Lipid Metabolism	13	2

K00864	E2.7.1.30, glpK	glycerol kinase [EC:2.7.1.30]	Environmental Adaptation - Endocrine System - Lipid Metabolism	3	3
K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	Endocrine System - Transport and Catabolism - Lipid Metabolism	4	2
ko03410_Base excision repair					
K01971	E6.5.1.1, lig	DNA ligase (ATP) [EC:6.5.1.1]	Replication and Repair	71	5
K01972	E6.5.1.2, ligA, ligB	DNA ligase (NAD+) [EC:6.5.1.2]	Replication and Repair	10	2
K02335	DPO1, polA	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	921	8
K03575	mutY	A/G-specific adenine glycosylase [EC:3.2.2.-]	Replication and Repair	4	2
K03648	UNG, UDG	uracil-DNA glycosylase [EC:3.2.2.-]	Immune System Diseases - Replication and Repair	119	3
K07462	recJ	single-stranded-DNA-specific exonuclease [EC:3.1.-.-]	Replication and Repair	4	2
K10563	mutM, fpg	formamidopyrimidine-DNA glycosylase [EC:3.2.2.23 4.2.99.18]	Replication and Repair	2	1
K10747	LIG1	DNA ligase 1 [EC:6.5.1.1]	Replication and Repair	1	1
K10773	NTH	endonuclease III [EC:4.2.99.18]	Replication and Repair	3	2
ko03420_Nucleotide excision repair					
K01971	E6.5.1.1, lig	DNA ligase (ATP) [EC:6.5.1.1]	Replication and Repair	71	5
K01972	E6.5.1.2, ligA, ligB	DNA ligase (NAD+) [EC:6.5.1.2]	Replication and Repair	10	2
K02335	DPO1, polA	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	921	8
K03657	uvrD, pcrA	DNA helicase II / ATP-dependent DNA helicase PcrA [EC:3.6.4.12]	Replication and Repair	110	3
K03701	uvrA	excinuclease ABC subunit A	Replication and Repair	5	2
K03702	uvrB	excinuclease ABC subunit B	Replication and Repair	5	2
K03703	uvrC	excinuclease ABC subunit C	Replication and Repair	78	2
K03723	mfd	transcription-repair coupling factor (superfamily II helicase) [EC:3.6.4.-]	Replication and Repair	2	1
K10747	LIG1	DNA ligase 1 [EC:6.5.1.1]	Replication and Repair	1	1
K10843	ERCC3, XPB	DNA excision repair protein ERCC-3 [EC:3.6.4.12]	Transcription - Replication and Repair	4	2
ko03430_Mismatch repair					
K01141	E3.1.11.1, sbcB	exodeoxyribonuclease I [EC:3.1.11.1]	Replication and Repair	1	1
K01971	E6.5.1.1, lig	DNA ligase (ATP) [EC:6.5.1.1]	Replication and Repair	71	5
K01972	E6.5.1.2, ligA, ligB	DNA ligase (NAD+) [EC:6.5.1.2]	Replication and Repair	10	2
K02337	DPO3A1, dnaE	DNA polymerase III subunit alpha [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	178	7
K02338	DPO3B, dnaN	DNA polymerase III subunit beta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	18	2
K02340	DPO3D1, holA	DNA polymerase III subunit delta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02341	DPO3D2, holB	DNA polymerase III subunit delta' [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02342	DPO3E, dnaQ	DNA polymerase III subunit epsilon [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	16	3
K02343	DPO3G, dnaX	DNA polymerase III subunit gamma/tau [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	13	3
K03111	ssb	single-strand DNA-binding protein	Replication and Repair	222	8
K03555	mutS	DNA mismatch repair protein MutS	Replication and Repair	3	3
K03572	mutL	DNA mismatch repair protein MutL	Replication and Repair	3	1
K03601	xseA	exodeoxyribonuclease VII large subunit [EC:3.1.11.6]	Replication and Repair	1	1
K03657	uvrD, pcrA	DNA helicase II / ATP-dependent DNA helicase PcrA [EC:3.6.4.12]	Replication and Repair	110	3
K03763	DPO3A2, polC	DNA polymerase III subunit alpha, Gram-positive type [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	9	3
K06223	dam	DNA adenine methylase [EC:2.1.1.72]	Replication and Repair	127	8
K07456	mutS2	DNA mismatch repair protein MutS2	Replication and Repair	1	1
K07462	recJ	single-stranded-DNA-specific exonuclease [EC:3.1.-.-]	Replication and Repair	4	2
K10747	LIG1	DNA ligase 1 [EC:6.5.1.1]	Replication and Repair	1	1
K10857	exoX	exodeoxyribonuclease X [EC:3.1.11.-]	Replication and Repair	1	1
ko03440_Homologous recombination					
K01159	ruvC	crossover junction endodeoxyribonuclease RuvC [EC:3.1.22.4]	Replication and Repair	76	2
K02335	DPO1, polA	DNA polymerase I [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	921	8
K02337	DPO3A1, dnaE	DNA polymerase III subunit alpha [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	178	7
K02338	DPO3B, dnaN	DNA polymerase III subunit beta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	18	2
K02340	DPO3D1, holA	DNA polymerase III subunit delta [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02341	DPO3D2, holB	DNA polymerase III subunit delta' [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	1	1
K02342	DPO3E, dnaQ	DNA polymerase III subunit epsilon [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	16	3
K02343	DPO3G, dnaX	DNA polymerase III subunit gamma/tau [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	13	3
K03111	ssb	single-strand DNA-binding protein	Replication and Repair	222	8
K03551	ruvB	holliday junction DNA helicase RuvB	Replication and Repair	4	2
K03553	recA	recombination protein RecA	Replication and Repair	335	6
K03581	recD	exodeoxyribonuclease V alpha subunit [EC:3.1.11.5]	Replication and Repair	6	1
K03582	recB	exodeoxyribonuclease V beta subunit [EC:3.1.11.5]	Replication and Repair	2	2
K03629	recF	DNA replication and repair protein RecF	Replication and Repair	1	1
K03655	recG	ATP-dependent DNA helicase RecG [EC:3.6.4.12]	Replication and Repair	11	3
K03763	DPO3A2, polC	DNA polymerase III subunit alpha, Gram-positive type [EC:2.7.7.7]	Nucleotide Metabolism - Replication and Repair	9	3
K04066	priA	primosomal protein N' (replication factor Y) (superfamily II helicase) [EC:3.6.4.-]	Replication and Repair	1	1
K07462	recJ	single-stranded-DNA-specific exonuclease [EC:3.1.-.-]	Replication and Repair	4	2
ko03450_Non-homologous end-joining					
K01971	E6.5.1.1, lig	DNA ligase (ATP) [EC:6.5.1.1]	Replication and Repair	71	5
ko04011_MAPK signaling pathway - yeast					
K01759	E4.4.1.5, GLO1, gloA	lactoylglutathione lyase [EC:4.4.1.5]	Carbohydrate Metabolism - Signal Transduction	8	1
ko04070_Phosphatidylinositol signaling system					
K01092	E3.1.3.25, IMPA, suhB	myo-inositol-1(or 4)-monophosphatase [EC:3.1.3.25]	Carbohydrate Metabolism - Signal Transduction - Biosynthesis of Other Secondary Metabolites	3	2
ko04112_Cell cycle - Caulobacter					
K01338	E3.4.21.53, lon	ATP-dependent Lon protease [EC:3.4.21.53]	Enzyme Families - Cell Growth and Death	5	3
K01358	clpP, CLPP	ATP-dependent Clp protease, protease subunit [EC:3.4.21.92]	Enzyme Families - Cell Growth and Death	163	6
K02313	dnaA	chromosomal replication initiator protein	Signal Transduction - Cell Growth and Death - Replication and Repair	470	2
K02314	dnaB	replicative DNA helicase [EC:3.6.4.12]	Cell Growth and Death - Replication and Repair	1499	9
K03544	clpX, CLPX	ATP-dependent Clp protease ATP-binding subunit ClpX	Folding, Sorting and Degradation - Cell Growth and Death	25	6
K03588	ftsW, spoVE	cell division protein FtsW	Cell Growth and Death - Replication and Repair	1	1
K03590	ftsA	cell division protein FtsA	Cell Growth and Death - Replication and Repair	1	1
K13581	ccrM	modification methylase [EC:2.1.1.72]	Cell Growth and Death	18	6
ko04113_Meiosis - yeast					
K01768	E4.6.1.1	adenylate cyclase [EC:4.6.1.1]	Nucleotide Metabolism - Cell Growth and Death	61	5
ko04115_p53 signaling pathway					
K10808	RRM2	ribonucleoside-diphosphate reductase subunit M2 [EC:1.17.4.1]	Metabolism of Other Amino Acids - Nucleotide Metabolism - Cell Growth and Death - Replication and Repair	67	5
ko04122_Sulfur relay system					
K00566	mnmA, trmU, TRMU	tRNA-specific 2-thiouridylase [EC:2.8.1.-]	Folding, Sorting and Degradation	4	3
K03154	thiS	sulfur carrier protein	Folding, Sorting and Degradation	1	1
K03639	MOCS1, moaA	molybdenum cofactor biosynthesis protein	Folding, Sorting and Degradation - Metabolism of Cofactors and Vitamins	1	1
K04487	iscS, NFS1	cysteine desulfurase [EC:2.8.1.7]	Folding, Sorting and Degradation - Metabolism of Cofactors and Vitamins	4	3
K11179	tusE, dsrC	tRNA 2-thiouridine synthesizing protein E [EC:2.8.1.-]	Folding, Sorting and Degradation	1	1
ko04141_Protein processing in endoplasmic reticulum					
K04079	htpG, HSP90A	molecular chaperone HtpG	Cancers - Environmental Adaptation - Immune System - Folding, Sorting and Degradation - Endocrine System	2	2
K13993	HSP20	HSP20 family protein	Folding, Sorting and Degradation	10	2

ko04142_Lysosome					
K12373	HEXA_B	hexosaminidase [EC:3.2.1.52]	Carbohydrate Metabolism - Folding, Sorting and Degradation - Glycan Biosynthesis and Metabolism - Transport and Catabolism	1	1
ko04146_Peroxisome					
K00273	E1.4.3.3, DAO	D-amino-acid oxidase [EC:1.4.3.3]	Metabolism of Other Amino Acids - Biosynthesis of Other Secondary Metabolites - Transport and Catabolism - Amino Acid Metabolism	2	1
K00869	E2.7.1.36, MVK, mvaK1	mevalonate kinase [EC:2.7.1.36]	Transport and Catabolism - Metabolism of Terpenoids and Polyketides	1	1
K01640	E4.1.3.4, HMGCL, hmgL	hydroxymethylglutaryl-CoA lyase [EC:4.1.3.4]	Carbohydrate Metabolism - Transport and Catabolism - Metabolism of Terpenoids and Polyketides - Amino Acid Metabolism - Lipid Metabolism	18	1
K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	Endocrine System - Transport and Catabolism - Lipid Metabolism	4	2
K03781	katE, CAT	catalase [EC:1.11.1.6]	Neurodegenerative Diseases - Transport and Catabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K04564	E1.15.1.1A, sodA, sodB, SOD2	superoxide dismutase, Fe-Mn family [EC:1.15.1.1]	Neurodegenerative Diseases - Transport and Catabolism	126	5
K04565	E1.15.1.1C, sodC, SOD1	Cu/Zn superoxide dismutase [EC:1.15.1.1]	Neurodegenerative Diseases - Transport and Catabolism	1	1
ko04210_Apoptosis					
K01173	E3.1.30.-	endonuclease [EC:3.1.30.-]	Cell Growth and Death	9	2
ko04260_Cardiac muscle contraction					
K00412	CYTB, petB	ubiquinol-cytochrome c reductase cytochrome b subunit [EC:1.10.2.2]	Circulatory System - Neurodegenerative Diseases - Energy Metabolism	7	2
ko04510_Focal adhesion					
K06236	COL1AS	collagen, type I/II/III/V/XI, alpha	Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	15	3
K06237	COL4A	collagen, type IV, alpha	Cancers - Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	7	2
ko04512_ECM-receptor interaction					
K06236	COL1AS	collagen, type I/II/III/V/XI, alpha	Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	15	3
K06237	COL4A	collagen, type IV, alpha	Cancers - Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	7	2
ko04610_Complement and coagulation cascades					
K01344	PROC	protein C (activated) [EC:3.4.21.69]	Immune System - Enzyme Families	1	1
ko04612_Antigen processing and presentation					
K04079	htpG, HSP90A	molecular chaperone HtpG	Cancers - Environmental Adaptation - Immune System - Folding, Sorting and Degradation - Endocrine System	2	2
ko04621_NOD-like receptor signaling pathway					
K04079	htpG, HSP90A	molecular chaperone HtpG	Cancers - Environmental Adaptation - Immune System - Folding, Sorting and Degradation - Endocrine System	2	2
ko04626_Plant-pathogen interaction					
K00864	E2.7.1.30, glpK	glycerol kinase [EC:2.7.1.30]	Environmental Adaptation - Endocrine System - Lipid Metabolism	3	3
K02358	EF-TU, tufA	elongation factor EF-Tu [EC:3.6.5.3]	Environmental Adaptation - Translation	2	2
K04079	htpG, HSP90A	molecular chaperone HtpG	Cancers - Environmental Adaptation - Immune System - Folding, Sorting and Degradation - Endocrine System	2	2
K13472	raxST	sulfotransferase	Environmental Adaptation	199	5
ko04724_Glutamatergic synapse					
K01915	E6.3.1.2, glnA	glutamine synthetase [EC:6.3.1.2]	Signal Transduction - Nervous System - Amino Acid Metabolism - Energy Metabolism	113	7
ko04910_Insulin signaling pathway					
K00688	E2.4.1.1, glgP, PYG	starch phosphorylase [EC:2.4.1.1]	Carbohydrate Metabolism - Endocrine System	9	3
ko04914_Progesterone-mediated oocyte maturation					
K04079	htpG, HSP90A	molecular chaperone HtpG	Cancers - Environmental Adaptation - Immune System - Folding, Sorting and Degradation - Endocrine System	2	2
ko04920_Adipocytokine signaling pathway					
K01897	ACSL, fadD	long-chain acyl-CoA synthetase [EC:6.2.1.3]	Endocrine System - Transport and Catabolism - Lipid Metabolism	4	2
ko04940_Type I diabetes mellitus					
K04077	groEL, HSPD1	chaperonin GroEL	Metabolic Diseases - Folding, Sorting and Degradation - Infectious Diseases	1783	8
ko04964_Proximal tubule bicarbonate reclamation					
K00261	E1.4.1.3	glutamate dehydrogenase (NAD(P)+) [EC:1.4.1.3]	Metabolism of Other Amino Acids - Excretory System - Amino Acid Metabolism - Energy Metabolism	3	2
ko04973_Carbohydrate digestion and absorption					
K01176	E3.2.1.1, amyA, malS	alpha-amylase [EC:3.2.1.1]	Carbohydrate Metabolism - Digestive System	2	1
ko04974_Protein digestion and absorption					
K06236	COL1AS	collagen, type I/II/III/V/XI, alpha	Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	15	3
K06237	COL4A	collagen, type IV, alpha	Cancers - Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	7	2
ko04978_Mineral absorption					
K00510	HMOX, hmuO, ho	heme oxygenase [EC:1.14.99.3]	Digestive System - Metabolism of Cofactors and Vitamins	2	1
ko05010_Alzheimer's disease					
K00134	GAPDH, gapA	glyceraldehyde 3-phosphate dehydrogenase [EC:1.2.1.12]	Carbohydrate Metabolism - Neurodegenerative Diseases	3	2
K00412	CYTB, petB	ubiquinol-cytochrome c reductase cytochrome b subunit [EC:1.10.2.2]	Circulatory System - Neurodegenerative Diseases - Energy Metabolism	7	2
ko05012_Parkinson's disease					
K00412	CYTB, petB	ubiquinol-cytochrome c reductase cytochrome b subunit [EC:1.10.2.2]	Circulatory System - Neurodegenerative Diseases - Energy Metabolism	7	2
ko05014_Amyotrophic lateral sclerosis (ALS)					
K03781	katE, CAT	catalase [EC:1.11.1.6]	Neurodegenerative Diseases - Transport and Catabolism - Amino Acid Metabolism - Energy Metabolism	1	1
K04565	E1.15.1.1C, sodC, SOD1	Cu/Zn superoxide dismutase [EC:1.15.1.1]	Neurodegenerative Diseases - Transport and Catabolism	1	1
ko05016_Huntington's disease					
K00412	CYTB, petB	ubiquinol-cytochrome c reductase cytochrome b subunit [EC:1.10.2.2]	Circulatory System - Neurodegenerative Diseases - Energy Metabolism	7	2
K03120	TBP, tbp	transcription initiation factor TFIID TATA-box-binding protein	Transcription - Neurodegenerative Diseases	3	1

K04564	E1.15.1.1A, sodA, sodB, SOD2	superoxide dismutase, Fe-Mn family [EC:1.15.1.1]	Neurodegenerative Diseases - Transport and Catabolism	126	5
K04565	E1.15.1.1C, sodC, SOD1	Cu/Zn superoxide dismutase [EC:1.15.1.1]	Neurodegenerative Diseases - Transport and Catabolism	1	1
ko05020_Prion diseases					
K04565	E1.15.1.1C, sodC, SOD1	Cu/Zn superoxide dismutase [EC:1.15.1.1]	Neurodegenerative Diseases - Transport and Catabolism	1	1
ko05100_Bacterial invasion of epithelial cells					
K13730	inlA	internalin A	Infectious Diseases	2	2
K13732	fnbA	fibronectin-binding protein A	Infectious Diseases	2	2
K13733	fnbB	fibronectin-binding protein B	Infectious Diseases	1	1
ko05110_Vibrio cholerae infection					
K08604	E3.4.24.25	vibriolysin [EC:3.4.24.25]	Enzyme Families - Infectious Diseases	1	1
ko05111_Vibrio cholerae pathogenic cycle					
K02405	fliA	RNA polymerase sigma factor for flagellar operon FliA	Cell Motility - Transcription - Signal Transduction - Infectious Diseases	7	2
K03087	SIG2, rpoS	RNA polymerase nonessential primary-like sigma factor	Transcription - Infectious Diseases	30	3
K03092	SIG54, rpoN	RNA polymerase sigma-54 factor	Transcription - Signal Transduction - Infectious Diseases	8	1
K07173	luxS	S-ribosylhomocysteine lyase [EC:4.4.1.21]	Infectious Diseases - Amino Acid Metabolism	1	1
K08604	E3.4.24.25	vibriolysin [EC:3.4.24.25]	Enzyme Families - Infectious Diseases	1	1
K08720	ompU	outer membrane protein OmpU	Infectious Diseases	2	1
ko05120_Epithelial cell signaling in Helicobacter pylori infection					
K01428	ureC	urease subunit alpha [EC:3.5.1.5]	Infectious Diseases - Nucleotide Metabolism - Xenobiotics Biodegradation and Metabolism - Amino Acid Metabolism	2	1
ko05142_Chagas disease (American trypanosomiasis)					
K01354	ptrB	oligopeptidase B [EC:3.4.21.83]	Enzyme Families - Infectious Diseases	2	2
ko05143_African trypanosomiasis					
K01354	ptrB	oligopeptidase B [EC:3.4.21.83]	Enzyme Families - Infectious Diseases	2	2
ko05146_Amoebiasis					
K01476	E3.5.3.1, rocF, arg	arginase [EC:3.5.3.1]	Infectious Diseases - Amino Acid Metabolism	11	1
K06236	COL1AS	collagen, type I/II/III/V/XI, alpha	Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	15	3
K06237	COL4A	collagen, type IV, alpha	Cancers - Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	7	2
ko05150_Staphylococcus aureus infection					
K14195	sasG	surface protein G	Infectious Diseases	8	2
K14205	mprF, fmtC	phosphatidylglycerol lysyltransferase [EC:2.3.2.3]	Infectious Diseases	3	1
ko05152_Tuberculosis					
K02040	pstS	phosphate transport system substrate-binding protein	Signal Transduction - Infectious Diseases - Membrane Transport	57	4
K04043	dnaK	molecular chaperone DnaK	Folding, Sorting and Degradation - Infectious Diseases	109	6
K04077	groEL, HSPD1	chaperonin GroEL	Metabolic Diseases - Folding, Sorting and Degradation - Infectious Diseases	1783	8
K14952	namH	UDP-MurNAc hydroxylase	Infectious Diseases	1	1
ko05200_Pathways in cancer					
K04079	htpG, HSP90A	molecular chaperone HtpG	Cancers - Environmental Adaptation - Immune System - Folding, Sorting and Degradation - Endocrine System	2	2
K06237	COL4A	collagen, type IV, alpha	Cancers - Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	7	2
ko05215_Prostate cancer					
K04079	htpG, HSP90A	molecular chaperone HtpG	Cancers - Environmental Adaptation - Immune System - Folding, Sorting and Degradation - Endocrine System	2	2
ko05222_Small cell lung cancer					
K06237	COL4A	collagen, type IV, alpha	Cancers - Digestive System - Signaling Molecules and Interaction - Infectious Diseases - Cell Communication	7	2
ko05322_Systemic lupus erythematosus					
K11089	TROVE2, SSA2	60 kDa SS-A/Ro ribonucleoprotein	Immune System Diseases	5	4
ko05340_Primary immunodeficiency					
K03648	UNG, UDG	uracil-DNA glycosylase [EC:3.2.2.-]	Immune System Diseases - Replication and Repair	119	3
no pathway					
K00046	idnO	gluconate 5-dehydrogenase [EC:1.1.1.69]	Metabolism	1	1
K00096	E1.1.1.261	glycerol-1-phosphate dehydrogenase [NAD(P)] [EC:1.1.1.261]	Metabolism	1	1
K00098	idnD	L-idonate 5-dehydrogenase [EC:1.1.1.264]	Metabolism	1	1
K00153	E1.1.1.306	S-(hydroxymethyl)mycothiol dehydrogenase [EC:1.1.1.306]	Metabolism	1	1
K00344	E1.6.5.5, qor	NADPH2:quinone reductase [EC:1.6.5.5]	Metabolism	2	2
K00346	nqrA	Na+-transporting NADH:ubiquinone oxidoreductase subunit A [EC:1.6.5.-]	Metabolism	1	1
K00355	E1.6.5.2, NQO1	NAD(P)H dehydrogenase (quinone) [EC:1.6.5.2]	Metabolism	1	1
K00428	E1.11.1.5	cytochrome c peroxidase [EC:1.11.1.5]	Cellular Processes and Signaling	2	2
K00435	E1.11.1.-	peroxiredoxin [EC:1.11.1.-]	Genetic Information Processing	6	4
K00518	E1.15.1.1	superoxide dismutase [EC:1.15.1.1]	Cellular Processes and Signaling	2	1
K00528	E1.18.1.2, fpr	ferredoxin--NADP+ reductase [EC:1.18.1.2]	Metabolism	1	1
K00540	E1.-.-.-		Metabolism	5	2
K00561	ermC, ermA	23S rRNA (adenine2085-N6)-dimethyltransferase [EC:2.1.1.184]	Genetic Information Processing	1	1
K00567	E2.1.1.63, MGMT, ogt	methylated-DNA-[protein]-cysteine S-methyltransferase [EC:2.1.1.63]	Replication and Repair	2	1
K00571	E2.1.1.72	site-specific DNA-methyltransferase (adenine-specific) [EC:2.1.1.72]	Genetic Information Processing	212	8
K00573	E2.1.1.77, pcm	protein-L-isoaspartate(D-aspartate) O-methyltransferase [EC:2.1.1.77]	Genetic Information Processing	1	1
K00590	E2.1.1.113	site-specific DNA-methyltransferase (cytosine-N4-specific) [EC:2.1.1.113]	Genetic Information Processing	52	4
K00612	E2.1.3.-	carbamoyltransferase [EC:2.1.3.-]	Genetic Information Processing	221	7
K00633	E2.3.1.18, lacA	galactoside O-acetyltransferase [EC:2.3.1.18]	Metabolism	18	3
K00638	E2.3.1.28, cat	chloramphenicol O-acetyltransferase [EC:2.3.1.28]	Metabolism	14	3
K00661	E2.3.1.79, maa	maltose O-acetyltransferase [EC:2.3.1.79]	Metabolism	6	4
K00666	K00666	fatty-acyl-CoA synthase [EC:6.2.1.-]	Lipid Metabolism	1	1
K00685	ATE1, ate1	arginine-tRNA-protein transferase [EC:2.3.2.8]	Genetic Information Processing	1	1
K00809	E2.5.1.46, dys1	deoxyhypusine synthase [EC:2.5.1.46]	Genetic Information Processing	3	3
K00837	E2.6.1.-		Metabolism	4	3
K00870	E2.7.1.37	protein kinase [EC:2.7.1.37]	Cellular Processes and Signaling	1	1
K00906	aceK	isocitrate dehydrogenase kinase/phosphatase [EC:2.7.11.5 3.1.3.-]	Cellular Processes and Signaling	2	1
K00924	E2.7.1.-		Cellular Processes and Signaling	1	1
K00960	E2.7.7.6	DNA-directed RNA polymerase [EC:2.7.7.6]	Genetic Information Processing	2	1
K00961	E2.7.7.7	DNA polymerase [EC:2.7.7.7]	Genetic Information Processing	4	2
K00996	E2.7.8.6, rfbP	undecaprenyl-phosphate galactose phosphotransferase [EC:2.7.8.6]	Glycan Biosynthesis and Metabolism	1	1
K01043	E2.-.-.-		Metabolism	1	1
K01090	E3.1.3.16	protein phosphatase [EC:3.1.3.16]	Metabolism	4	1

K01104	E3.1.3.48	protein-tyrosine phosphatase [EC:3.1.3.48]	Cellular Processes and Signaling	2	1
K01133	E3.1.6.6, betC	choline-sulfatase [EC:3.1.6.6]	Cellular Processes and Signaling	2	2
K01138	K01138	uncharacterized sulfatase [EC:3.1.6.-]	Cellular Processes and Signaling	1	1
K01143	E3.1.11.3	exodeoxyribonuclease (lambda-induced) [EC:3.1.11.3]	Genetic Information Processing	6	3
K01144	E3.1.11.5	exodeoxyribonuclease V [EC:3.1.11.5]	Genetic Information Processing	3	2
K01153	hsdR	type I restriction enzyme, R subunit [EC:3.1.21.3]	Genetic Information Processing	67	3
K01154	hsdS	type I restriction enzyme, S subunit [EC:3.1.21.3]	Genetic Information Processing	15	5
K01155	E3.1.21.4	type II restriction enzyme [EC:3.1.21.4]	Genetic Information Processing	12	3
K01156	res	type III restriction enzyme [EC:3.1.21.5]	Genetic Information Processing	11	2
K01160	rusA	crossover junction endodeoxyribonuclease RusA [EC:3.1.22.4]	Replication and Repair	20	3
K01161	E3.1.25.1	deoxyribonuclease (pyrimidine dimer) [EC:3.1.25.1]	Metabolism	1	1
K01174	E3.1.31.1, nuc	micrococcal nuclease [EC:3.1.31.1]	Genetic Information Processing	13	2
K01175	E3.1.-.-		Metabolism	2	2
K01181	E3.2.1.8, xynA	endo-1,4-beta-xylanase [EC:3.2.1.8]	Metabolism	21	2
K01185	E3.2.1.17	lysozyme [EC:3.2.1.17]	Metabolism	3297	7
K01200	E3.2.1.41	pullulanase [EC:3.2.1.41]	Metabolism	1	1
K01219	E3.2.1.81	agarase [EC:3.2.1.81]	Metabolism	1	1
K01236	E3.2.1.141	maltooligosyltrehalose trehalohydrolase [EC:3.2.1.141]	Metabolism	1	1
K01238	E3.2.1.-		Metabolism	26	2
K01265	E3.4.11.18, map	methionyl aminopeptidase [EC:3.4.11.18]	Enzyme Families	4	2
K01269	E3.4.11.-	aminopeptidase [EC:3.4.11.-]	Metabolism	1	1
K01284	dcp	peptidyl-dipeptidase Dcp [EC:3.4.15.5]	Enzyme Families	1	1
K01297	ldcA	muramoyltetrapeptide carboxypeptidase [EC:3.4.17.13]	Enzyme Families	1	1
K01299	E3.4.17.19	carboxypeptidase Taq [EC:3.4.17.19]	Enzyme Families	3	2
K01308	E3.4.19.11	g-D-glutamyl-meso-diaminopimelate peptidase [EC:3.4.19.11]	Enzyme Families	1	1
K01317	ACR	acrosin [EC:3.4.21.10]	Enzyme Families	1	1
K01322	E3.4.21.26, PREP	prolyl oligopeptidase [EC:3.4.21.26]	Enzyme Families	1	1
K01356	lexA	repressor LexA [EC:3.4.21.88]	Enzyme Families - Replication and Repair	8	5
K01361	E3.4.21.96	lactocepin [EC:3.4.21.96]	Folding, Sorting and Degradation - Enzyme Families	2	2
K01362	E3.4.21.-		Metabolism	30	2
K01407	ptr	protease III [EC:3.4.24.55]	Enzyme Families	1	1
K01409	E3.4.24.57, gcp	O-sialoglycoprotein endopeptidase [EC:3.4.24.57]	Enzyme Families	1	1
K01417	E3.4.24.-		Metabolism	3	2
K01419	hslV, clpQ	ATP-dependent HslUV protease, peptidase subunit HslV [EC:3.4.25.-]	Enzyme Families	1	1
K01422	E3.4.99.-		Cellular Processes and Signaling	1	1
K01446	E3.5.1.28	N-acetylmuramoyl-L-alanine amidase [EC:3.5.1.28]	Cellular Processes and Signaling	10	1
K01447	E3.5.1.28A, cwlA, xlyA, xlyB	N-acetylmuramoyl-L-alanine amidase [EC:3.5.1.28]	Cellular Processes and Signaling	159	4
K01448	E3.5.1.28B, amiA, amiB, amiC	N-acetylmuramoyl-L-alanine amidase [EC:3.5.1.28]	Replication and Repair	257	3
K01449	E3.5.1.28C, cwlJ, sleB	N-acetylmuramoyl-L-alanine amidase [EC:3.5.1.28]	Cellular Processes and Signaling	3	2
K01462	PDF, def	peptide deformylase [EC:3.5.1.88]	Metabolism	23	6
K01529	E3.6.1.-		Metabolism	1	1
K01533	E3.6.3.4, ATP7, copA	Cu2+-exporting ATPase [EC:3.6.3.4]	Metabolism	3	2
K01537	E3.6.3.8	Ca2+-transporting ATPase [EC:3.6.3.8]	Metabolism	2	2
K01551	E3.6.3.16, arsA	arsenite-transporting ATPase [EC:3.6.3.16]	Cellular Processes and Signaling	4	1
K01669	E4.1.99.3, phrB	deoxyribodipyrimidine photo-lyase [EC:4.1.99.3]	Replication and Repair	1	1
K01724	E4.2.1.96, PCBD, phhB	4a-hydroxytetrahydrobiopterin dehydratase [EC:4.2.1.96]	Metabolism	7	2
K01795	E5.1.3.-		Metabolism	20	6
K01802	E5.2.1.8	peptidylprolyl isomerase [EC:5.2.1.8]	Genetic Information Processing	10	3
K01854	glf	UDP-galactopyranose mutase [EC:5.4.99.9]	Cellular Processes and Signaling	123	5
K01989	ABC.X4.S	putative ABC transport system substrate-binding protein	Membrane Transport	1	1
K01990	ABC-2.A	ABC-2 type transport system ATP-binding protein	Membrane Transport	9	2
K01991	ABC-2.OM, wza	polysaccharide export outer membrane protein	Cellular Processes and Signaling	2	2
K02003	ABC.CD.A		Membrane Transport	3	2
K02004	ABC.CD.P		Membrane Transport	4	2
K02014	ABC.FEV.OM	iron complex outermembrane recepter protein	Cellular Processes and Signaling	6	3
K02023	ABC.MS.A	multiple sugar transport system ATP-binding protein	Membrane Transport	1	1
K02025	ABC.MS.P	multiple sugar transport system permease protein	Membrane Transport	2	2
K02026	ABC.MS.P1	multiple sugar transport system permease protein	Membrane Transport	3	3
K02027	ABC.MS.S	multiple sugar transport system substrate-binding protein	Membrane Transport	3	1
K02055	ABC.SP.S	putative spermidine/putrescine transport system substrate-binding protein	Membrane Transport	1	1
K02058	ABC.SS.S	simple sugar transport system substrate-binding protein	Membrane Transport	4	2
K02065	ABC.X1.A	putative ABC transport system ATP-binding protein	Membrane Transport	1	1
K02066	ABC.X1.P	putative ABC transport system permease protein	Membrane Transport	1	1
K02078	acpP	acyl carrier protein	Metabolism	2	1
K02198	ccmF	cytochrome c-type biogenesis protein CcmF	Cellular Processes and Signaling	1	1
K02238	comEC	competence protein ComEC	Membrane Transport	2	2
K02283	cpaF, tadA	pilus assembly protein CpaF	Cell Motility - Membrane Transport	1	1
K02315	dnaC	DNA replication protein DnaC	Replication and Repair	20	1
K02334	dpo	DNA polymerase bacteriophage-type [EC:2.7.7.7]	Genetic Information Processing	190	7
K02336	DPO2, polB	DNA polymerase II [EC:2.7.7.7]	Replication and Repair	27	2
K02346	DPO4, dinB	DNA polymerase IV [EC:2.7.7.7]	Replication and Repair	1	1
K02355	EF-G, fusA	elongation factor EF-G [EC:3.6.5.3]	Translation	2	2
K02395	flgJ	flagellar protein FlgJ	Cell Motility	47	3
K02404	flhF	flagellar biosynthesis protein FlhF	Cell Motility	1	1
K02437	gcvH	glycine cleavage system H protein	Metabolism	1	1
K02440	GLPF	glycerol uptake facilitator protein	Signaling Molecules and Interaction	1	1
K02469	gyrA	DNA gyrase subunit A [EC:5.99.1.3]	Replication and Repair	18	3
K02470	gyrB	DNA gyrase subunit B [EC:5.99.1.3]	Replication and Repair	15	3
K02481	K02481	two-component system, NtrC family, response regulator	Signal Transduction	1	1
K02493	hemK	methyltransferase [EC:2.1.1.-]	Genetic Information Processing	3	2
K02503	hit	Hit-like protein involved in cell-cycle regulation	Poorly Characterized	1	1
K02518	IF-1, infA	translation initiation factor IF-1	Translation	4	4
K02519	IF-2, infB	translation initiation factor IF-2	Translation	3	2
K02520	IF-3, infC	translation initiation factor IF-3	Translation	6	2
K02528	ksgA	16S rRNA (adenine1518-N6/adenine1519-N6)-dimethyltransferase [EC:2.1.1.182]	Translation	2	1
K02598	nirC	nitrite transporter NirC	Cellular Processes and Signaling	2	1
K02601	nusG	transcriptional antiterminator NusG	Translation	6	2
K02609	paaA	phenylacetic acid degradation protein	Poorly Characterized	11	2
K02622	parE	topoisomerase IV subunit B [EC:5.99.1.-]	Replication and Repair	4	4
K02666	pilQ	type IV pilus assembly protein PilQ	Cell Motility - Membrane Transport	2	1
K02674	pilY1	type IV pilus assembly protein PilY1	Cell Motility - Membrane Transport	2	1
K02687	prmA	ribosomal protein L11 methyltransferase [EC:2.1.1.-]	Genetic Information Processing	1	1
K02715	psbN	PsbN protein	Poorly Characterized	2	1
K02742	sprT	SprT protein	Poorly Characterized	1	1
K02805	rffA, wecE	lipopolysaccharide biosynthesis protein	Amino Acid Metabolism	5	3
K02835	RF-1, prfA	peptide chain release factor RF-1	Translation	1	1
K02837	RF-3, prfC	peptide chain release factor RF-3	Translation	1	1
K03168	topA	DNA topoisomerase I [EC:5.99.1.2]	Replication and Repair	10	3
K03207	wcaH	colanic acid biosynthesis protein WcaH [EC:3.6.1.-]	Metabolism	3	3
K03281	TC.CIC	chloride channel protein, CIC family	Cellular Processes and Signaling	2	1
K03284	TC.MIT	metal ion transporter, MIT family	Cellular Processes and Signaling	8	3
K03286	TC.OOP	OmpA-OmpF porin, OOP family	Cellular Processes and Signaling	3	2
K03292	TC.GPH	glycoside/pentoside/hexuronide:cation symporter, GPH family	Cellular Processes and Signaling	1	1
K03293	TC.AAT	amino acid transporter, AAT family	Cellular Processes and Signaling	1	1
K03294	TC.APA	basic amino acid/polyamine antiporter, APA family	Cellular Processes and Signaling	1	1
K03295	TC.CDF	cation efflux system protein, CDF family	Cellular Processes and Signaling	1	1
K03296	TC.HAE1	hydrophobic/amphiphilic exporter-1 (mainly G- bacteria), HAE1 family	Cellular Processes and Signaling	1	1
K03298	TC.DME	drug/metabolite transporter, DME family	Cellular Processes and Signaling	1	1
K03299	TC.GNTP	gluconate:H+ symporter, GntP family	Cellular Processes and Signaling	1	1
K03303	TC.LCTP	lactate transporter, LctP family	Cellular Processes and Signaling	1	1

K03307	TC.SSS	solute:Na+ symporter, SSS family	Cellular Processes and Signaling	1	1
K03320	TC.AMT	ammonium transporter, Amt family	Cellular Processes and Signaling	2	2
K03328	TC.PST	polysaccharide transporter, PST family	Cellular Processes and Signaling	1	1
K03386	E1.11.1.15, PRDX, ahpC	peroxiredoxin (alkyl hydroperoxide reductase subunit C) [EC:1.11.1.15]	Genetic Information Processing	6	5
K03427	hsdM	type I restriction enzyme M protein [EC:2.1.1.72]	Genetic Information Processing	29	5
K03438	mraW	S-adenosyl-methyltransferase [EC:2.1.1.-]	Cellular Processes and Signaling	2	2
K03439	E2.1.1.33	tRNA (guanine-N7-)-methyltransferase [EC:2.1.1.33]	Metabolism	7	2
K03457	TC.NCS1	nucleobase:cation symporter-1, NCS1 family	Cellular Processes and Signaling	4	1
K03466	ftsK, spoIIIE	DNA segregation ATPase FtsK/SpoIIIE, S-DNA-T family	Replication and Repair	15	3
K03495	gidA	glucose inhibited division protein A	Replication and Repair	1	1
K03496	parA, soj	chromosome partitioning protein	Cell Motility - Replication and Repair	80	5
K03497	parB, spo0J	chromosome partitioning protein, ParB family	Cell Motility - Transcription - Replication and Repair	25	7
K03502	DPO5C, umuC	DNA polymerase V	Replication and Repair	6	2
K03503	DPO5D, umuD	DNA polymerase V [EC:3.4.21.-]	Enzyme Families - Replication and Repair	4	2
K03521	fixA, etfB	electron transfer flavoprotein beta subunit	Metabolism	1	1
K03522	fixB, etfA	electron transfer flavoprotein alpha subunit	Metabolism	2	2
K03523	bioY	putative biotin biosynthesis protein BioY	Poorly Characterized	1	1
K03529	smc	chromosome segregation protein	Replication and Repair	7	1
K03530	hupB	DNA-binding protein HU-beta	Replication and Repair	3	2
K03545	tig	trigger factor	Genetic Information Processing	9	4
K03546	sbcC	exonuclease SbcC	Replication and Repair	55	5
K03547	sbcD	exonuclease SbcD	Replication and Repair	4	2
K03561	exbB	biopolymer transport protein ExbB	Cellular Processes and Signaling	1	1
K03574	MUTT, NUDT1, MTH1	7,8-dihydro-8-oxoguanine triphosphatase [EC:3.6.1.-]	Replication and Repair	12	2
K03578	hrpA	ATP-dependent helicase HrpA [EC:3.6.4.13]	Genetic Information Processing	1	1
K03580	hepA	ATP-dependent helicase HepA [EC:3.6.4.-]	Genetic Information Processing	6	1
K03585	acrA	membrane fusion protein	Replication and Repair	1	1
K03596	lepA	GTP-binding protein LepA	Cellular Processes and Signaling	2	2
K03606	wcaJ	putative colanic acid biosynthesis UDP-glucose lipid carrier transferase	Cellular Processes and Signaling	1	1
K03624	greA	transcription elongation factor GreA	Genetic Information Processing	1	1
K03625	nusB	N utilization substance protein B	Translation	1	1
K03630	radC	DNA repair protein RadC	Genetic Information Processing	21	5
K03641	tolB	TolB protein	Cellular Processes and Signaling	1	1
K03642	rlpA	rare lipoprotein A	Cellular Processes and Signaling	87	2
K03643	lptE, rlpB	LPS-assembly lipoprotein	Cellular Processes and Signaling	1	1
K03646	tolA	colicin import membrane protein	Cellular Processes and Signaling	5	2
K03656	rep	ATP-dependent DNA helicase Rep [EC:3.6.4.12]	Replication and Repair	1	1
K03664	smpB	SsrA-binding protein	Genetic Information Processing	40	2
K03671	trxA	thioredoxin 1	Folding, Sorting and Degradation	136	4
K03672	trxC	thioredoxin 2 [EC:1.8.1.8]	Folding, Sorting and Degradation	10	2
K03676	grxC, GLRX, GLRX2	glutaredoxin 3	Folding, Sorting and Degradation	42	6
K03684	rnd	ribonuclease D [EC:3.1.13.5]	Genetic Information Processing	5	2
K03686	dnaJ	molecular chaperone DnaJ	Folding, Sorting and Degradation	65	4
K03694	clpA	ATP-dependent Clp protease ATP-binding subunit ClpA	Folding, Sorting and Degradation	26	5
K03695	clpB	ATP-dependent Clp protease ATP-binding subunit ClpB	Folding, Sorting and Degradation	45	6
K03696	clpC	ATP-dependent Clp protease ATP-binding subunit ClpC	Folding, Sorting and Degradation	342	8
K03697	clpE	ATP-dependent Clp protease ATP-binding subunit ClpE	Folding, Sorting and Degradation	7	3
K03704	cspA	cold shock protein (beta-ribbon, CspA family)	Transcription	2	2
K03711	fur	Fur family transcriptional regulator, ferric uptake regulator	Transcription	2	1
K03716	splB	spore photoproduct lyase [EC:4.1.99.-]	Genetic Information Processing	6	2
K03721	tyrR	transcriptional regulator of aroF, aroG, tyrA and aromatic amino acid transport	Transcription	3	1
K03724	lhr	ATP-dependent helicase Lhr and Lhr-like helicase [EC:3.6.4.-]	Replication and Repair	1	1
K03726	helS	helicase [EC:3.6.4.-]	Genetic Information Processing	3	1
K03727	helY	ATP-dependent RNA helicase HelY [EC:3.6.4.-]	Genetic Information Processing	1	1
K03733	xerC	integrase/recombinase XerC	Replication and Repair	25	2
K03743	K03743		Poorly Characterized	4	4
K03744	lemA	LemA protein	Poorly Characterized	1	1
K03750	moeA	molybdopterin biosynthesis protein MoeA	Metabolism	3	2
K03762	proP	MFS transporter, MHS family, proline/betaine transporter	Membrane Transport	1	1
K03767	PPIA	peptidyl-prolyl cis-trans isomerase A (cyclophilin A) [EC:5.2.1.8]	Folding, Sorting and Degradation	2	2
K03768	PPIB, ppiB	peptidyl-prolyl cis-trans isomerase B (cyclophilin B) [EC:5.2.1.8]	Folding, Sorting and Degradation	24	2
K03771	surA	peptidyl-prolyl cis-trans isomerase SurA [EC:5.2.1.8]	Folding, Sorting and Degradation	1	1
K03774	slpA	FKBP-type peptidyl-prolyl cis-trans isomerase SlpA [EC:5.2.1.8]	Folding, Sorting and Degradation	7	3
K03775	slyD	FKBP-type peptidyl-prolyl cis-trans isomerase SlyD [EC:5.2.1.8]	Folding, Sorting and Degradation	7	1
K03789	rimI	ribosomal-protein-alanine N-acetyltransferase [EC:2.3.1.128]	Translation	3	2
K03791	K03791	putative chitinase	Poorly Characterized	1162	6
K03796	bax	Bax protein	Poorly Characterized	45	3
K03797	E3.4.21.102, prc, ctpA	carboxyl-terminal processing protease [EC:3.4.21.102]	Enzyme Families	1	1
K03798	ftsH, hflB	cell division protease FtsH [EC:3.4.24.-]	Folding, Sorting and Degradation - Enzyme Families	3	2
K03802	cphA	cyanophycin synthetase [EC:6.-.-.-]	Cellular Processes and Signaling	4	1
K03806	ampD	AmpD protein	Cellular Processes and Signaling	2	1
K03818	wcaF	putative colanic acid biosynthesis acetyltransferase WcaF [EC:2.3.1.-]	Metabolism	1	1
K03832	tonB	periplasmic protein TonB	Cellular Processes and Signaling	3	2
K03839	fldA	flavodoxin I	Metabolism	6	1
K03840	fldB	flavodoxin II	Metabolism	1	1
K03924	moxR	MoxR-like ATPase [EC:3.6.3.-]	Metabolism	16	4
K03977	engA	GTP-binding protein	Translation	3	1
K03979	obg	GTP-binding protein	Translation	2	1
K04024	eutJ	ethanolamine utilization protein EutJ	Metabolism	1	1
K04047	dps	starvation-inducible DNA-binding protein	Replication and Repair	66	2
K04068	nrdG	anaerobic ribonucleoside-triphosphate reductase activating protein [EC:1.97.1.4]	Genetic Information Processing	4	2
K04075	tilS, mesJ	tRNA(Ile)-lysidine synthase [EC:6.3.4.-]	Genetic Information Processing	2	2
K04078	groES, HSPE1	chaperonin GroES	Folding, Sorting and Degradation	308	4
K04080	ibpA	molecular chaperone IbpA	Folding, Sorting and Degradation	78	6
K04090	E1.2.7.8	indolepyruvate ferredoxin oxidoreductase [EC:1.2.7.8]	Metabolism	1	1
K04091	ssuD	alkanesulfonate monooxygenase [EC:1.14.14.5]	Metabolism	3	2
K04095	fic	cell filamentation protein	Replication and Repair	1	1
K04096	smf	DNA processing protein	Genetic Information Processing	1	1
K04338	csgF	curli production assembly/transport component CsgF	Membrane Transport	1	1
K04483	radA	DNA repair protein RadA	Replication and Repair	1	1
K04485	sms, radA	DNA repair protein RadA/Sms	Replication and Repair	110	2
K04488	iscU, nifU	nitrogen fixation protein NifU and related proteins	Metabolism	97	5
K04651	hypA	hydrogenase nickel incorporation protein HypA	Genetic Information Processing	1	1
K04652	hypB	hydrogenase nickel incorporation protein HypB	Genetic Information Processing	1	1
K04654	hypD	hydrogenase expression/formation protein HypD	Genetic Information Processing	5	1
K04744	lptD, imp, ostA	LPS-assembly protein	Cellular Processes and Signaling	1	1
K04750	phnB	PhnB protein	Poorly Characterized	1	1
K04756	ahpD	alkyl hydroperoxide reductase subunit D	Poorly Characterized	1	1
K04758	feoA	ferrous iron transport protein A	Cellular Processes and Signaling	1	1
K04759	feoB	ferrous iron transport protein B	Cellular Processes and Signaling	1	1
K04763	xerD	integrase/recombinase XerD	Replication and Repair	58	6
K04768	acuC	acetoin utilization protein AcuC	Metabolism	4	2
K04773	sppA	protease IV [EC:3.4.21.-]	Enzyme Families	10	2
K04800	rfcL	replication factor C large subunit	Replication and Repair	3	2
K04801	rfcS	replication factor C small subunit	Replication and Repair	456	7
K05303	E2.1.1.101	macrocin O-methyltransferase [EC:2.1.1.101]	Metabolism	7	2
K05337	fer	ferredoxin	Metabolism	9	3
K05516	cbpA	curved DNA-binding protein	Folding, Sorting and Degradation - Replication and Repair	27	5
K05520	pfpI	protease I [EC:3.2.-.-]	Enzyme Families	10	2
K05521	draG	ADP-ribosylglycohydrolase [EC:3.2.-.-]	Genetic Information Processing	8	2
K05524	fdxA	ferredoxin	Metabolism	12	7

K05548	benK	MFS transporter, AAHS family, benzoate transport protein	Membrane Transport	1	1
K05559	phaA	multicomponent K+:H+ antiporter subunit A	Cellular Processes and Signaling	1	1
K05786	rarD	chloramphenicol-sensitive protein RarD	Cellular Processes and Signaling	1	1
K05795	terD	tellurium resistance protein TerD	Cellular Processes and Signaling	1	1
K05807	comL	putative lipoprotein	Poorly Characterized	1	1
K05838	ybbN	putative thioredoxin	Folding, Sorting and Degradation	4	2
K05844	rimK	ribosomal protein S6 modification protein	Translation	7	2
K05989	E3.2.1.40	alpha-L-rhamnosidase [EC:3.2.1.40]	Metabolism	1	1
K06013	E3.4.24.84	STE24 endopeptidase [EC:3.4.24.84]	Enzyme Families	1	1
K06024	scpB	segregation and condensation protein B	Replication and Repair	1	1
K06041	E5.3.1.13	arabinose-5-phosphate isomerase [EC:5.3.1.13]	Cellular Processes and Signaling	7	3
K06048	ybdK	carboxylate-amine ligase [EC:6.3.-.-]	Metabolism	1	1
K06135	pqqA	pyrroloquinoline quinone biosynthesis protein A	Poorly Characterized	9	2
K06143	creD	inner membrane protein	Cellular Processes and Signaling	1	1
K06147	ABCB-BAC	ATP-binding cassette, subfamily B, bacterial	Membrane Transport	5	2
K06158	ABCF3	ATP-binding cassette, sub-family F, member 3	Genetic Information Processing	1	1
K06168	miaB	bifunctional enzyme involved in thiolation and methylation of tRNA	Genetic Information Processing	1	1
K06177	rluA	ribosomal large subunit pseudouridine synthase A [EC:5.4.99.12]	Translation	1	1
K06178	rluB	ribosomal large subunit pseudouridine synthase B [EC:5.4.99.12]	Translation	1	1
K06180	rluD	ribosomal large subunit pseudouridine synthase D [EC:5.4.99.12]	Translation	1	1
K06189	corC	magnesium and cobalt transporter	Cellular Processes and Signaling	1	1
K06192	pqiB	paraquat-inducible protein B	Poorly Characterized	2	1
K06204	dksA	DnaK suppressor protein	Translation - Transcription	3	2
K06214	csgG	curli production assembly/transport component CsgG	Membrane Transport	23	6
K06217	phoH, phoL	phosphate starvation-inducible protein PhoH and related proteins	Cellular Processes and Signaling	475	7
K06351	kipI	inhibitor of KinA	Cellular Processes and Signaling	1	1
K06400	spoIVCA	site-specific DNA recombinase	Cellular Processes and Signaling	6	3
K06405	spoVAC	stage V sporulation protein AC	Cellular Processes and Signaling	1	1
K06415	spoVR	stage V sporulation protein R	Cellular Processes and Signaling	20	4
K06416	spoVS	stage V sporulation protein S	Cellular Processes and Signaling	8	2
K06860	K06860		Poorly Characterized	1	1
K06872	K06872		Poorly Characterized	1	1
K06877	K06877		Poorly Characterized	2	1
K06879	queF	7-cyano-7-deazaguanine reductase [EC:1.7.1.13]	Poorly Characterized	49	4
K06881	K06881		Poorly Characterized	1	1
K06884	K06884		Poorly Characterized	11	1
K06885	K06885		Poorly Characterized	1	1
K06890	K06890		Poorly Characterized	9	3
K06891	clpS	ATP-dependent Clp protease adaptor protein ClpS	Genetic Information Processing	31	4
K06894	K06894		Poorly Characterized	1	1
K06901	pbuG	putative MFS transporter, AGZA family, xanthine/uracil permease	Membrane Transport	4	3
K06902	UMF1	MFS transporter, UMF1 family	Membrane Transport	1	1
K06903	K06903		Poorly Characterized	131	3
K06904	K06904		Poorly Characterized	496	6
K06905	K06905		Poorly Characterized	4	1
K06906	K06906		Poorly Characterized	4	1
K06907	K06907		Poorly Characterized	627	7
K06909	K06909		Poorly Characterized	1744	8
K06911	K06911		Poorly Characterized	1	1
K06915	K06915		Poorly Characterized	11	3
K06917	selU	tRNA 2-selenouridine synthase [EC:2.9.1.-]	Poorly Characterized	1	1
K06918	K06918		Poorly Characterized	1	1
K06919	K06919	putative DNA primase/helicase	Genetic Information Processing	716	8
K06920	queC	queuosine biosynthesis protein QueC	Poorly Characterized	396	8
K06927	K06927		Poorly Characterized	2	1
K06938	K06938		Poorly Characterized	5	3
K06940	K06940		Poorly Characterized	7	2
K06941	rlmN	ribosomal RNA large subunit methyltransferase N [EC:2.1.1.-]	Translation	2	1
K06950	K06950	uncharacterized protein	Poorly Characterized	1	1
K06952	K06952		Poorly Characterized	1	1
K06955	K06955		Poorly Characterized	1	1
K06960	K06960		Poorly Characterized	1	1
K06966	K06966		Poorly Characterized	2	2
K06969	rlmI	ribosomal RNA large subunit methyltransferase I [EC:2.1.1.-]	Translation	2	1
K06972	K06972		Poorly Characterized	1	1
K06980	K06980		Poorly Characterized	1	1
K06991	K06991		Poorly Characterized	1	1
K06994	K06994	putative drug exporter of the RND superfamily	Poorly Characterized	2	1
K06995	K06995		Poorly Characterized	3	2
K07000	K07000		Poorly Characterized	1	1
K07001	K07001		Poorly Characterized	1	1
K07003	K07003		Poorly Characterized	2	2
K07004	K07004		Poorly Characterized	1	1
K07008	DUG3	glutamine amidotransferase	Poorly Characterized	12	2
K07010	K07010	putative glutamine amidotransferase	Enzyme Families	4	2
K07011	K07011		Poorly Characterized	49	2
K07012	K07012		Poorly Characterized	2	1
K07019	K07019		Poorly Characterized	2	1
K07024	K07024		Poorly Characterized	1	1
K07031	K07031		Poorly Characterized	15	5
K07033	K07033		Poorly Characterized	1	1
K07043	K07043		Poorly Characterized	1	1
K07044	K07044		Poorly Characterized	1	1
K07052	K07052		Poorly Characterized	1	1
K07053	K07053		Poorly Characterized	1	1
K07056	K07056		Poorly Characterized	1	1
K07058	K07058	membrane protein	Poorly Characterized	2	1
K07067	disA	DNA integrity scanning protein	Poorly Characterized	1	1
K07071	K07071		Poorly Characterized	2	1
K07074	K07074		Poorly Characterized	1	1
K07088	K07088		Poorly Characterized	2	2
K07090	K07090		Poorly Characterized	1	1
K07093	K07093		Poorly Characterized	1	1
K07098	K07098		Poorly Characterized	1	1
K07099	K07099		Poorly Characterized	1	1
K07100	K07100		Poorly Characterized	1	1
K07101	K07101		Poorly Characterized	40	2
K07103	K07103		Poorly Characterized	1	1
K07107	ybgC	acyl-CoA thioester hydrolase [EC:3.1.2.-]	Poorly Characterized	1	1
K07114	K07114	uncharacterized protein	Poorly Characterized	2	2
K07117	K07117		Poorly Characterized	4	3
K07124	K07124		Poorly Characterized	1	1
K07126	K07126		Poorly Characterized	6	2
K07130	K07130		Poorly Characterized	1	1
K07137	K07137		Poorly Characterized	271	4
K07154	K07154		Poorly Characterized	3	2
K07156	pcoC		Poorly Characterized	1	1
K07168	K07168	CBS domain-containing membrane protein	Cellular Processes and Signaling	4	1
K07169	K07169	FHA domain-containing protein	Cellular Processes and Signaling	1	1
K07175	phoH2	PhoH-like ATPase	Cellular Processes and Signaling	186	6
K07180	prkA	serine protein kinase	Cellular Processes and Signaling	52	5
K07182	K07182	CBS domain-containing protein	Cellular Processes and Signaling	2	2
K07219	K07219	putative molybdopterin biosynthesis protein	Cellular Processes and Signaling	3	1
K07228	K07228	TrkA domain protein	Cellular Processes and Signaling	1	1
K07233	copB	copper resistance protein B	Cellular Processes and Signaling	1	1
K07239	TC.HME	heavy-metal exporter, HME family	Cellular Processes and Signaling	7	2

K07243	FTR1	high-affinity iron transporter	Cellular Processes and Signaling	1	1
K07261	mepA	penicillin-insensitive murein endopeptidase [EC:3.4.24.-]	Enzyme Families	1	1
K07262	pbpG	D-alanyl-D-alanine endopeptidase (penicillin-binding protein 7) [EC:3.4.99.-]	Enzyme Families	25	2
K07263	pqqL	zinc protease [EC:3.4.99.-]	Enzyme Families	4	1
K07270	K07270	glycosyl transferase, family 25	Cellular Processes and Signaling	21	3
K07272	rgpF	rhamnosyltransferase [EC:2.4.1.-]	Glycan Biosynthesis and Metabolism	1	1
K07273	acm	lysozyme	Cellular Processes and Signaling	2	1
K07288	tspA	uncharacterized membrane protein	Cellular Processes and Signaling	1	1
K07303	E1.3.99.16B	isoquinoline 1-oxidoreductase, beta subunit [EC:1.3.99.16]	Metabolism	1	1
K07304	msrA	peptide-methionine (S)-S-oxide reductase [EC:1.8.4.11]	Genetic Information Processing	1	1
K07316	mod	adenine-specific DNA-methyltransferase [EC:2.1.1.72]	Genetic Information Processing	3	2
K07317	K07317	adenine-specific DNA-methyltransferase [EC:2.1.1.72]	Genetic Information Processing	1	1
K07319	yhdJ	putative adenine-specific DNA-methyltransferase [EC:2.1.1.72]	Genetic Information Processing	780	9
K07335	bmpA, bmpB, tmpC	basic membrane protein A and related proteins	Cellular Processes and Signaling	1	1
K07336	K07336	PKHD-type hydroxylase [EC:1.14.11.-]	Metabolism	34	6
K07357	fimB	type 1 fimbriae regulatory protein FimB	Genetic Information Processing	3	2
K07358	fimE	type 1 fimbriae regulatory protein FimE	Genetic Information Processing	2	1
K07386	pepO	putative endopeptidase [EC:3.4.24.-]	Enzyme Families	1	1
K07389	cyaC, hlyC, rtxC	cytolysin-activating lysine-acyltransferase [EC:2.3.1.-]	Genetic Information Processing	5	1
K07391	comM	magnesium chelatase family protein	Genetic Information Processing	1	1
K07393	ECM4	putative glutathione S-transferase	Genetic Information Processing	1	1
K07394	K07394	SM-20-related protein	Genetic Information Processing	8	1
K07396	K07396	putative protein-disulfide isomerase	Genetic Information Processing	1	1
K07442	TRM61, GCD14	tRNA (adenine-N1-)-methyltransferase catalytic subunit [EC:2.1.1.36]	Genetic Information Processing	1	1
K07444	ypsC	putative N6-adenine-specific DNA methylase [EC:2.1.1.-]	Genetic Information Processing	1	1
K07452	K07452, mcrB	5-methylcytosine-specific restriction enzyme B [EC:3.1.21.-]	Genetic Information Processing	1	1
K07455	recT	recombination protein RecT	Replication and Repair	90	4
K07460	yraN	putative endonuclease	Genetic Information Processing	1	1
K07461	K07461	putative endonuclease	Genetic Information Processing	7	3
K07465	K07465	putative RecB family exonuclease	Genetic Information Processing	3	2
K07474	xtmA	phage terminase small subunit	Genetic Information Processing	24	5
K07478	ycaJ	putative ATPase	Genetic Information Processing	34	4
K07491	K07491	putative transposase	Genetic Information Processing	7	3
K07496	K07496	putative transposase	Genetic Information Processing	108	8
K07497	K07497	putative transposase	Genetic Information Processing	2	2
K07501	K07501	hypothetical protein	Genetic Information Processing	59	5
K07505	K07505	hypothetical protein	Genetic Information Processing	50	2
K07507	mgtC	putative Mg2+ transporter-C (MgtC) family protein	Cellular Processes and Signaling	2	2
K07552	bcr	MFS transporter, DHA1 family, bicyclomycin/chloramphenicol resistance protein	Membrane Transport	3	2
K07566	SUA5	putative translation factor	Translation	1	1
K07568	queA	S-adenosylmethionine:tRNA ribosyltransferase-isomerase [EC:5.-.-.-]	Genetic Information Processing	1	1
K07576	K07576	metallo-beta-lactamase family protein	Genetic Information Processing	1	1
K07589	folX	D-erythro-7,8-dihydroneopterin triphosphate epimerase [EC:5.-.-.-]	Metabolism	1	1
K07726	K07726	putative transcriptional regulator	Transcription	1	1
K07728	K07728	putative transcriptional regulator	Transcription	1	1
K07729	K07729	putative transcriptional regulator	Transcription	4	1
K07738	nrdR	transcriptional repressor NrdR	Transcription	3	1
K07741	K07741	anti-repressor protein	Genetic Information Processing	11	3
K07749	frc	formyl-CoA transferase [EC:2.8.3.16]	Metabolism	1	1
K08151	tetA	MFS transporter, DHA1 family, tetracycline resistance protein	Membrane Transport	1	1
K08160	cmr, mdfA	MFS transporter, DHA1 family, multidrug/chloramphenicol efflux transport protein	Membrane Transport	1	1
K08223	fsr	MFS transporter, FSR family, fosmidomycin resistance protein	Membrane Transport	1	1
K08259	lytM	lysostaphin [EC:3.4.24.75]	Enzyme Families	1	1
K08282	E2.7.11.1	non-specific serine/threonine protein kinase [EC:2.7.11.1]	Cellular Processes and Signaling	2	2
K08304	mltA	membrane-bound lytic murein transglycosylase A [EC:3.2.1.-]	Metabolism	1	1
K08307	mltD, dniR	membrane-bound lytic murein transglycosylase D [EC:3.2.1.-]	Metabolism	23	2
K08309	slt	soluble lytic murein transglycosylase [EC:3.2.1.-]	Metabolism	12	2
K08482	kaiC	circadian clock protein KaiC	Cellular Processes and Signaling	28	2
K08602	pepF, pepB	oligoendopeptidase F [EC:3.4.24.-]	Enzyme Families	1	1
K08884	K08884	serine/threonine protein kinase, bacterial [EC:2.7.11.1]	Enzyme Families	6	3
K08930	pucA	light-harvesting protein B-800-850 alpha chain	Energy Metabolism	1	1
K08981	K08981	putative membrane protein	Poorly Characterized	2	1
K08999	K08999	hypothetical protein	Poorly Characterized	1	1
K09005	K09005	hypothetical protein	Poorly Characterized	15	4
K09007	K09007	hypothetical protein	Poorly Characterized	12	5
K09013	sufC	Fe-S cluster assembly ATP-binding protein	Membrane Transport	3	2
K09014	sufB	Fe-S cluster assembly protein SufB	Poorly Characterized	13	3
K09117	K09117	hypothetical protein	Poorly Characterized	1	1
K09125	K09125	hypothetical protein	Poorly Characterized	4	3
K09129	K09129	hypothetical protein	Poorly Characterized	1	1
K09134	K09134	hypothetical protein	Poorly Characterized	6	1
K09136	ycaO	ribosomal protein S12 methylthiotransferase	Translation	1	1
K09139	K09139	hypothetical protein	Poorly Characterized	1	1
K09181	yfiQ	hypothetical protein	Poorly Characterized	1	1
K09607	ina	immune inhibitor A [EC:3.4.24.-]	Enzyme Families	1	1
K09728	K09728	hypothetical protein	Poorly Characterized	1	1
K09744	K09744	hypothetical protein	Poorly Characterized	8	1
K09746	K09746	hypothetical protein	Poorly Characterized	2	1
K09760	rmuC	DNA recombination protein RmuC	Genetic Information Processing	1	1
K09774	lptA	lipopolysaccharide export system protein LptA	Cellular Processes and Signaling	1	1
K09776	K09776	hypothetical protein	Poorly Characterized	10	3
K09786	K09786	hypothetical protein	Poorly Characterized	20	3
K09787	K09787	hypothetical protein	Poorly Characterized	1	1
K09790	K09790	hypothetical protein	Poorly Characterized	4	2
K09791	K09791	hypothetical protein	Poorly Characterized	1	1
K09818	ABC.MN.S	manganese/iron transport system substrate-binding protein	Membrane Transport	1	1
K09861	K09861	hypothetical protein	Poorly Characterized	2	1
K09892	K09892	hypothetical protein	Poorly Characterized	3	1
K09935	K09935	hypothetical protein	Poorly Characterized	3	1
K09946	K09946	hypothetical protein	Poorly Characterized	1	1
K09955	K09955	hypothetical protein	Poorly Characterized	2	1
K09960	K09960	hypothetical protein	Poorly Characterized	54	4
K09961	K09961	hypothetical protein	Poorly Characterized	26	2
K09966	K09966	hypothetical protein	Poorly Characterized	5	2
K09968	K09968	hypothetical protein	Poorly Characterized	2	2
K09973	K09973	hypothetical protein	Poorly Characterized	1	1
K10026	queE, ykvL, ygcF	queuosine biosynthesis protein QueE	Poorly Characterized	125	5
K10704	UBE2V	ubiquitin-conjugating enzyme E2 variant	Replication and Repair	1	1
K10726	mcm, cdc21	replicative DNA helicase Mcm [EC:3.6.4.-]	Replication and Repair	22	3
K10819	E2.7.13.3	histidine kinase	Metabolism	1	1
K10896	FANCM	fanconi anemia group M protein [EC:3.6.4.13]	Replication and Repair	2	1
K10906	recE	exodeoxyribonuclease VIII [EC:3.1.11.-]	Replication and Repair	33	2
K10908	POLRMT, RPO41	DNA-directed RNA polymerase, mitochondrial [EC:2.7.7.6]	Genetic Information Processing	36	4
K11068	hlyIII	hemolysin III	Signaling Molecules and Interaction	5	2
K11107	yfaE	ferredoxin	Metabolism	9	4
K11159	K11159	carotenoid cleavage dioxygenase	Metabolism	1	1
K11527	K11527	two-component system, unclassified family, sensor histidine kinase and response regulator [EC:2.7.13.3]	Signal Transduction - Enzyme Families	2	1
K11895	impH, vasB	type VI secretion system protein ImpH	Membrane Transport	2	1
K11900	impC	type VI secretion system protein ImpC	Membrane Transport	1	1
K12065	traB	conjugal transfer pilus assembly protein TraB	Membrane Transport	1	1
K12287	mshQ	MSHA biogenesis protein MshQ	Membrane Transport	1	1
K12507	fadK	acyl-CoA synthetase [EC:6.2.1.-]	Lipid Metabolism	1	1

K12684	K12684, sigA, sepA, esp	serine protease autotransporter [EC:3.4.21.-]	Membrane Transport	3	1
K12685	ssp	subtilase-type serine protease [EC:3.4.21.-]	Membrane Transport	1	1
K12950	ctpC	cation-transporting P-type ATPase C [EC:3.6.3.-]	Metabolism	1	1
K12979	lpxO	beta-hydroxylase [EC:1.14.11.-]	Glycan Biosynthesis and Metabolism	2	2
K12989	lpcC	mannosyltransferase [EC:2.4.1.-]	Glycan Biosynthesis and Metabolism	1	1
K12997	rgpB	rhamnosyltransferase [EC:2.4.1.-]	Glycan Biosynthesis and Metabolism	1	1
K13005	rfbV	abequosyltransferase [EC:2.4.1.-]	Glycan Biosynthesis and Metabolism	1	1
K13010	per	perosamine synthetase	Glycan Biosynthesis and Metabolism - Amino Acid Metabolism	14	4
K13013	wbqV	O-antigen biosynthesis protein WbqV	Glycan Biosynthesis and Metabolism	4	3
K13016	wbpB	UDP-D-GlcNAcA oxidase [EC:1.1.1.-]	Glycan Biosynthesis and Metabolism	8	2
K13018	wbpD	UDP-D-GlcNAc3NA acetyltransferase [EC:2.3.1.-]	Glycan Biosynthesis and Metabolism	1	1
K13019	wbpI	UDP-GlcNAc3NAcA epimerase [EC:5.1.3.23]	Glycan Biosynthesis and Metabolism	6	3
K13020	wlbA, bplA	UDP-D-GlcNAcA oxidase [EC:1.1.1.-]	Glycan Biosynthesis and Metabolism	2	2
K13057	treT	trehalose synthase [EC:2.4.1.245]	Metabolism	3	2
K13275	isp	major intracellular serine protease [EC:3.4.21.-]	Folding, Sorting and Degradation - Enzyme Families	1	1
K13281	uvsE, UVE1	UV DNA damage endonuclease [EC:3.-.-.-]	Genetic Information Processing	103	5
K13611	pksJ	polyketide synthase PksJ	Lipid Metabolism	11	1
K13628	iscA	iron-sulfur cluster assembly protein	Metabolism	54	3
K13635	cbl	LysR family transcriptional regulator, cys regulon transcriptional activator	Transcription	1	1
K13678	cpoA	monoglucosyldiacylglycerol glycosyltransferase [EC:2.4.1.-]	Glycan Biosynthesis and Metabolism	5	2
K13693	K13693	glucosyl-3-phosphoglycerate synthase [EC:2.4.1.-]	Glycan Biosynthesis and Metabolism	1	1
K13694	spr	lipoprotein Spr	Enzyme Families	7	2
K13695	nlpC	probable lipoprotein NlpC	Enzyme Families	8	2
K13819	K13819	NifU-like protein	Metabolism	2	1
K13930	mdcB	triphosphoribosyl-dephospho-CoA synthase [EC:2.7.8.25]	Metabolism	1	1
K14059	int	integrase	Genetic Information Processing	17	1
K14060	pinR	putative DNA-invertase from lambdoid prophage Rac	Genetic Information Processing	1	1
K14162	dnaE2	error-prone DNA polymerase [EC:2.7.7.7]	Replication and Repair	3	2
K14266	prnA	FADH2 O2-dependent halogenase I [EC:1.14.14.7]	Metabolism	25	4
K14287	ybdL	methionine aminotransferase [EC:2.6.1.-]	Amino Acid Metabolism	1	1
K14393	actP	cation/acetate symporter	Cellular Processes and Signaling	1	1
K14415	rtcB	protein RtcB	Poorly Characterized	1	1
K14441	rimO	ribosomal protein S12 methylthiotransferase [EC:2.-.-.-]	Translation	1	1
K14623	dinD	DNA-damage-inducible protein D	Poorly Characterized	42	1
K14645	K14645	serine protease [EC:3.4.21.-]	Folding, Sorting and Degradation - Enzyme Families	3	1
K14660	nodE	nodulation protein E [EC:2.3.1.-]	Cellular Processes and Signaling	1	1
K14665	amhX	amidohydrolase [EC:3.5.1.-]	Enzyme Families	1	1
K14680	E6.5.1.3	RNA ligase [EC:6.5.1.3]	Metabolism	72	3
K14744	rzpD	putative Rz endopeptidase from lambdoid prophage DLP12 [EC:3.4.-.-]	Poorly Characterized	8	2

Annexe A.5 : Matériel supplémentaire

Supplementary material : Assessing the diversity and specificity of two freshwater viral communities through metagenomics (Article IV)

	Lake Bourget	Lake Pavin
Coords	45°43'47" N, 5°52'10" E	45°29'41" N, 2°53'13" E
Trophic status	Mesotrophic	Oligomesotrophic
Depth	145m	92m
Length	18km	800m
Width	2.8km	800m
Area	44.5km²	0.44km²
Altitude	231.5m	1197m
Drainage basin	56 000 ha	50 ha

Table S1. Characteristics of the two lakes studied.

Virome	Environment	Extraction protocol	Number of sequences	Mean sequence size	BLAST hit ratio against NR for 100-bp reads
Lake Bourget	Freshwater	Peg	593 084	433	2.20%
Lake Pavin	Freshwater	Peg	684 224	412	0.70%
46 TilPondKentSTVir0806	Freshwater	Peg / CsCl	56 549	101	3.38%
47 TilPondKentSTVir050406	Freshwater	Peg / CsCl	60 135	101	1.91%
48 PrePondKentSTVir050406	Freshwater	Peg / CsCl	67 785	103	2.56%
49 TpondKentSTVir1105	Freshwater	Peg / CsCl	264 844	102	2.15%
Lake Limnopolar Spring	Freshwater	Sucrose	41 322	237	0.86%
Lake Limnopolar Summer	Freshwater	Sucrose	38 475	221	4.50%
60 pHPorCompHawVir0206	Eukaryote		49 949	105	5.38%
73 FishHealSlimKentSTVir050406	Eukaryote		61 022	99	14.72%
74 FishMorSlimKentSTVir050406	Eukaryote		59 599	98	23.00%
85 Mosquito ISDVir01252006	Eukaryote		336 760	103	28.60%
86 Mosquito DigSDVir060606	Eukaryote		638 689	101	35.44%
87 Mosquito IISDVir060606	Eukaryote		601 040	104	18.32%
13 MedSalternSDbayVir111605	Hypersaline	Peg / CsCl	58 319	98	3.10%
14 HighSalternSDbayVir111605	Hypersaline	Peg / CsCl	151 180	100	1.96%
15 LowSalternSDbayVir0704	Hypersaline	Peg / CsCl	268 049	104	2.58%
16 LowSalternSDbayVir111005	Hypersaline	Peg / CsCl	109 836	104	2.78%
18 MedSalternSDbayVir112205	Hypersaline	Peg / CsCl	55 142	101	1.62%
19 HighSalternSDbayVir120705	Hypersaline	Peg / CsCl	46 628	102	5.26%
21 LowSalternSDbayVir112805	Hypersaline	Peg / CsCl	62 363	104	18.14%
22 SaltonSeaVirOne082308	Hypersaline		55 467	104	1.66%
32 GOMVir94to01	Marine	CsCl	262 501	102	10.94%
33 BBCVir96to04	Marine	CsCl	414 964	102	4.99%
34 ArcticVir2002	Marine	CsCl	686 209	99	28.61%
35 SARVir063005	Marine	CsCl	397 939	104	4.46%
36 KingLIVir082105	Marine	CsCl	93 744	108	6.95%
37 XmasLIVir080505	Marine	CsCl	279 882	111	24.52%
38 PalmLIVir081805	Marine	CsCl	318 178	105	3.00%
39 FannLIVir081105	Marine	CsCl	378 475	104	2.82%

Table S2. Main characteristics of viromes included in the different comparisons. Extraction methodology is indicated where available. Peg = polyethylene glycol; CsCl = Cesium Chloride. The BLAST hit ratio is the percentage of reads significantly similar to a protein of the database the non-redundant database (threshold of 10^{-3} on e-value and 50 on scores). As sequence comparison results depend on the length of the sequences, reads were reduced to 100-bp long for all viromes.

	Lake Bourget virome	Lake Bourget bacterial metagenome	Lake Pavin virome	Lake Pavin 16S rRNA
Actinobacteria	6.79%	44.54%	3.85%	28.50%
Bacteroidetes	12.99%	9.21%	19.85%	14.30%
Chlamydiae	3.10%	0.00%	0.10%	0.00%
Cyanobacteria	6.40%	0.27%	5.08%	7.10%
Firmicutes	15.90%	1.75%	12.62%	0.00%
Alphaproteobacteria	17.00%	18.38%	11.54%	7.10%
Betaproteobacteria	9.89%	23.59%	14.31%	21.40%
Gammaproteobacteria	15.42%	0.17%	15.38%	0.00%
Deltaproteobacteria	4.75%	1.44%	6.62%	0.00%
Verrucomicrobia	1.36%	0.00%	2.31%	17.90%
Other *Bacteria*	6.40%	0.65%	8.34%	3.70%

Table S3. Bacterial taxonomic composition as deduced from virome reads best BLAST hits, compared with previously published data. These previous data are from a metagenome for Lake Bourget, and from 16SrRNA PCR amplification for Lake Pavin.

PFAM Id	PFAM name	PFAM description	Lake Bourget	Lake Pavin	Viral domain	Known in virus
PF02305	Phage_F	Capsid protein (F protein)	23197	856	X	X
PF09675	Chlamy_scaf	Chlamydia-phage Chp2 scaffold (Chlamy_scaf)	4784	78	X	X
PF01555	N6_N4_Mtase	DNA methylase	964	2450		X
PF02407	Viral_Rep	Putative viral replication protein	171	2858	X	X
PF00959	Phage_lysozyme	Phage lysozyme	495	2218	X	X
PF00145	DNA_methylase	C-5 cytosine-specific DNA methylase	1074	1437		X
PF04404	ERF	ERF superfamily	1171	1049		X
PF00118	Cpn60_TCP1	TCP-1/cpn60 chaperonin family	1183	245		X
PF03796	DnaB_C	DnaB-like helicase C terminal domain	1084	262		X
PF08291	Peptidase_M15_3	Peptidase M15	910	345		X
PF06378	DUF1071	Protein of unknown function (DUF1071)	711	523		X
PF04466	Terminase_3	Phage terminase large subunit	733	464	X	X
PF04860	Phage_portal	Phage portal protein	879	234	X	X
PF01370	Epimerase	NAD dependent epimerase/dehydratase family	768	273		X
PF02867	Ribonuc_red_lgC	Ribonucleotide reductase, barrel domain	815	200		X
PF07230	Phage_T4_Gp20	Bacteriophage T4-like capsid assembly protein (Gp20)	479	479	X	X
PF02511	Thy1	Thymidylate synthase complementing protein	787	127		X
PF03237	Terminase_6	Terminase-like family	538	351		X
PF08419	RNA_helicase	RNA helicase	10	836	X	X
PF04586	Peptidase_U35	Caudovirus prohead protease	609	217	X	X
PF07068	Gp23	Major capsid protein Gp23	643	145	X	X
PF08800	VirE_N	VirE N-terminal domain	344	382		X
PF05792	Candida_ALS	Candida agglutinin-like (ALS)	428	248		
PF01041	DegT_DnrJ_EryC1	DegT/DnrJ/EryC1/StrS aminotransferase family	493	133		X
PF05766	NinG	Bacteriophage Lambda NinG protein	247	379	X	X
PF00535	Glycos_transf_2	Glycosyl transferase family 2	346	269		X
PF04851	ResIII	Type III restriction enzyme, res subunit	216	379		X
PF01391	Collagen	Collagen triple helix repeat (20 copies)	346	247		X
PF01520	Amidase_3	N-acetylmuramoyl-L-alanine amidase	587	0		X
PF08299	Bac_DnaA_C	Bacterial dnaA protein helix-turn-helix / Replication initiation factor	196	303		X

GO Id	GO Name	Lake Bourget	Lake Pavin
GO:0019028	GO:viral capsid	23741	1031
GO:0005198	GO:structural molecule activity	23592	1041
GO:0003677	GO:DNA binding	5164	6264
GO:0006260	GO:DNA replication	3704	4618
GO:0005524	GO:ATP binding	4528	3001
GO:0006306	GO:DNA methylation	2142	3963
GO:0016779	GO:nucleotidyltransferase activity	239	3750
GO:0042624	GO:ATPase activity, uncoupled	181	3694
GO:0009253	GO:peptidoglycan catabolic process	1183	2347
GO:0008170	GO:N-methyltransferase activity	973	2460
GO:0016998	GO:cell wall macromolecule catabolic process	744	2296
GO:0018142	GO:protein-DNA covalent cross-linking	171	2858
GO:0016888	GO:endodeoxyribonuclease activity, producing 5'-phosphomonoesters	171	2858
GO:0003796	GO:lysozyme activity	501	2220
GO:0004519	GO:endonuclease activity	513	1103
GO:0003824	GO:catalytic activity	1121	443
GO:0003678	GO:DNA helicase activity	1248	272
GO:0005515	GO:protein binding	1185	246
GO:0044267	GO:cellular protein metabolic process	1183	245
GO:0006231	GO:dTMP biosynthetic process	1186	241
GO:0016787	GO:hydrolase activity	669	735
GO:0003887	GO:DNA-directed DNA polymerase activity	1068	303
GO:0004748	GO:ribonucleoside-diphosphate reductase activity	1055	314
GO:0003676	GO:nucleic acid binding	765	515
GO:0006323	GO:DNA packaging	760	478
GO:0046872	GO:metal ion binding	189	1020
GO:0050662	GO:coenzyme binding	771	273
GO:0044237	GO:cellular metabolic process	768	273
GO:0005971	GO:ribonucleoside-diphosphate reductase complex	815	200
GO:0050660	GO:FAD binding	127	787

Table S4. Main functions retrieved in the viromes. In the first table, the 30 most retrieved PFAM domains in the viromes are listed, with the number of sequences for each virome alongside informations about their description in viral genomes, or the fact that most of the sequences from this domain are of viral origin (identified as « viral » domains). In the second table, the 30 most retrieved GO terms are listed, with the associated number of sequences for each virome.

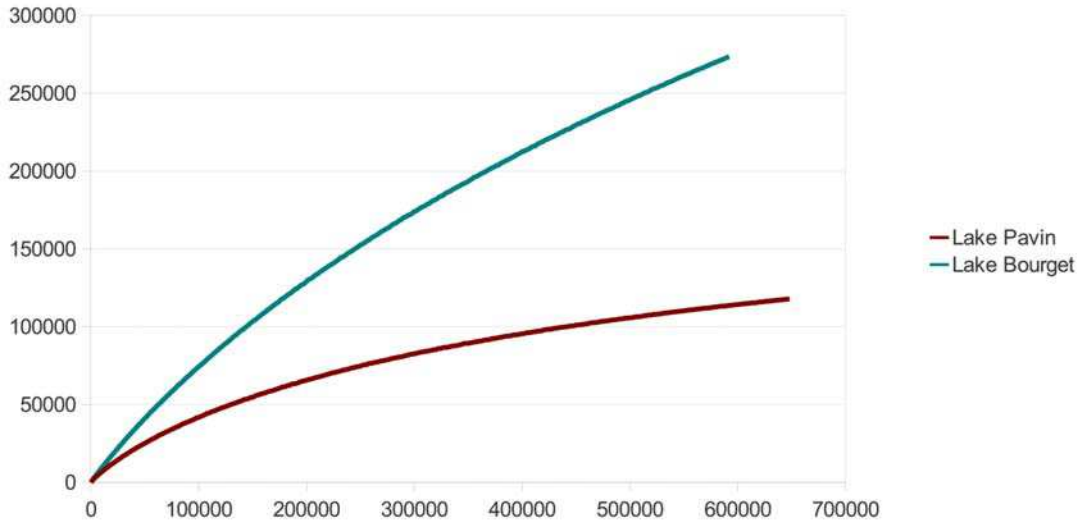


Figure S1. Rarefaction curves based on whole viromes. Each virome was clusterized at 75% identity, and the curve presents the number of different clusters as a function of the number of input sequences.

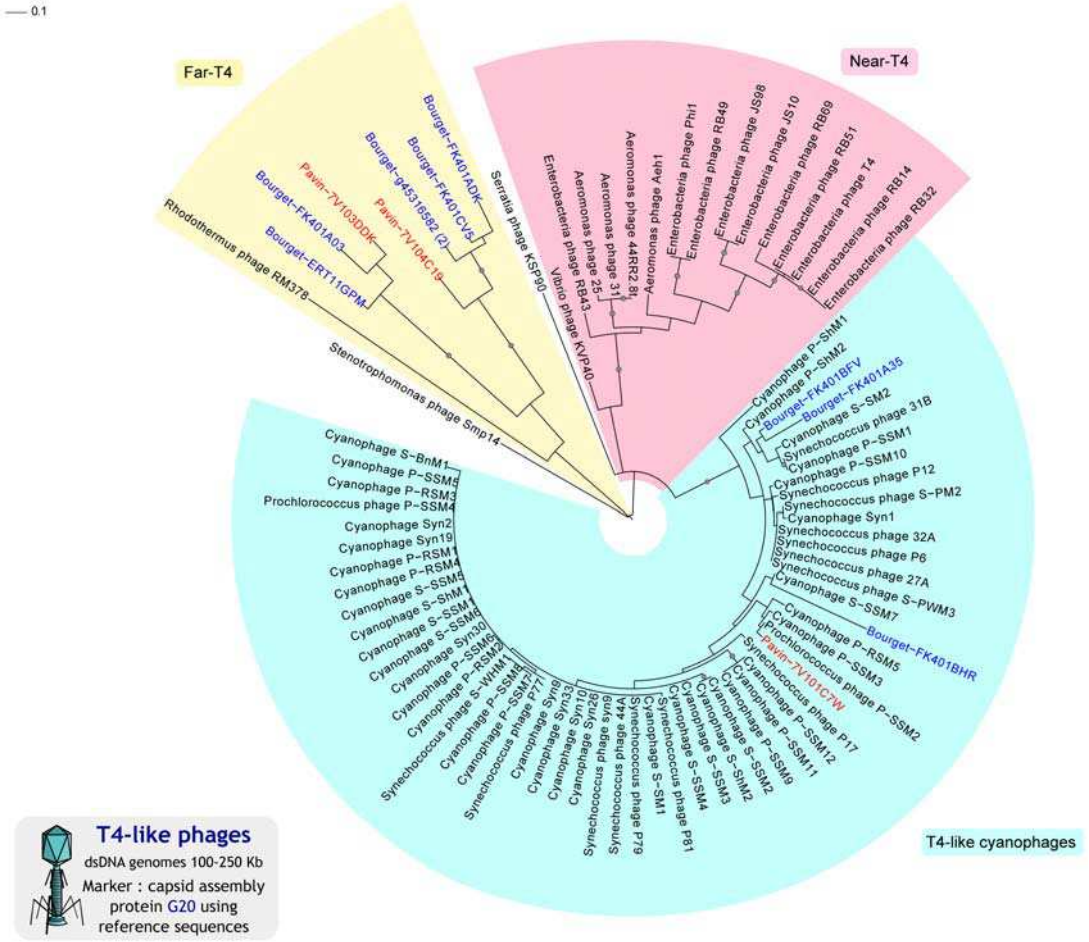


Figure S2. Maximum-likelihood tree for T4-like phages (G20). The main reference groups are indicated on the tree (near-T4 in red, T4-like cyanophages in blue), and the Far-T4 group is highlighted in yellow. Leaf labels corresponding to virome sequences are colored (red for Lake Pavin and blue for Lake Bourget). The number of reads assembled is given in brackets for each contig. Nodes with at least 80% bootstrap support are flagged with black circles.

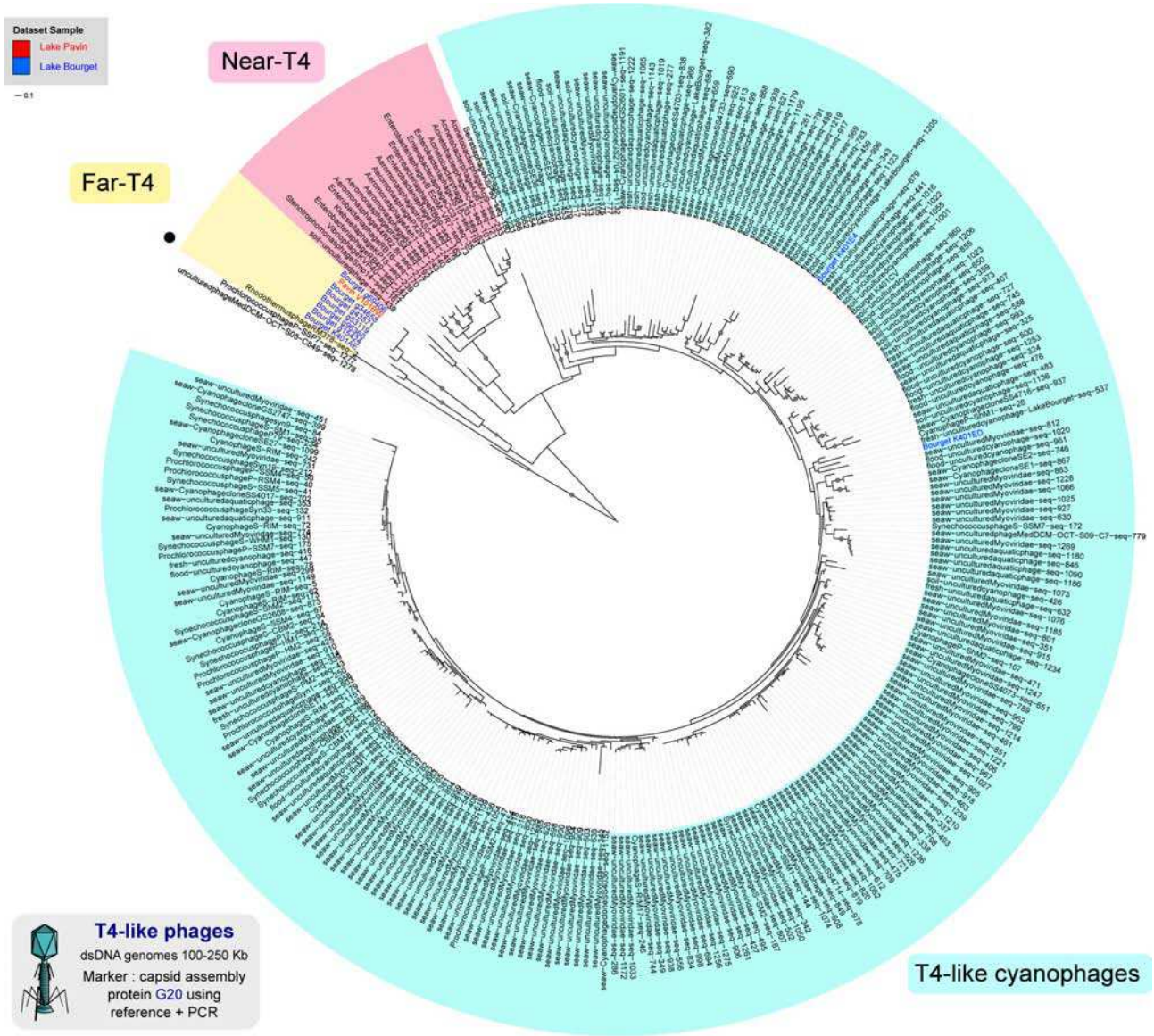
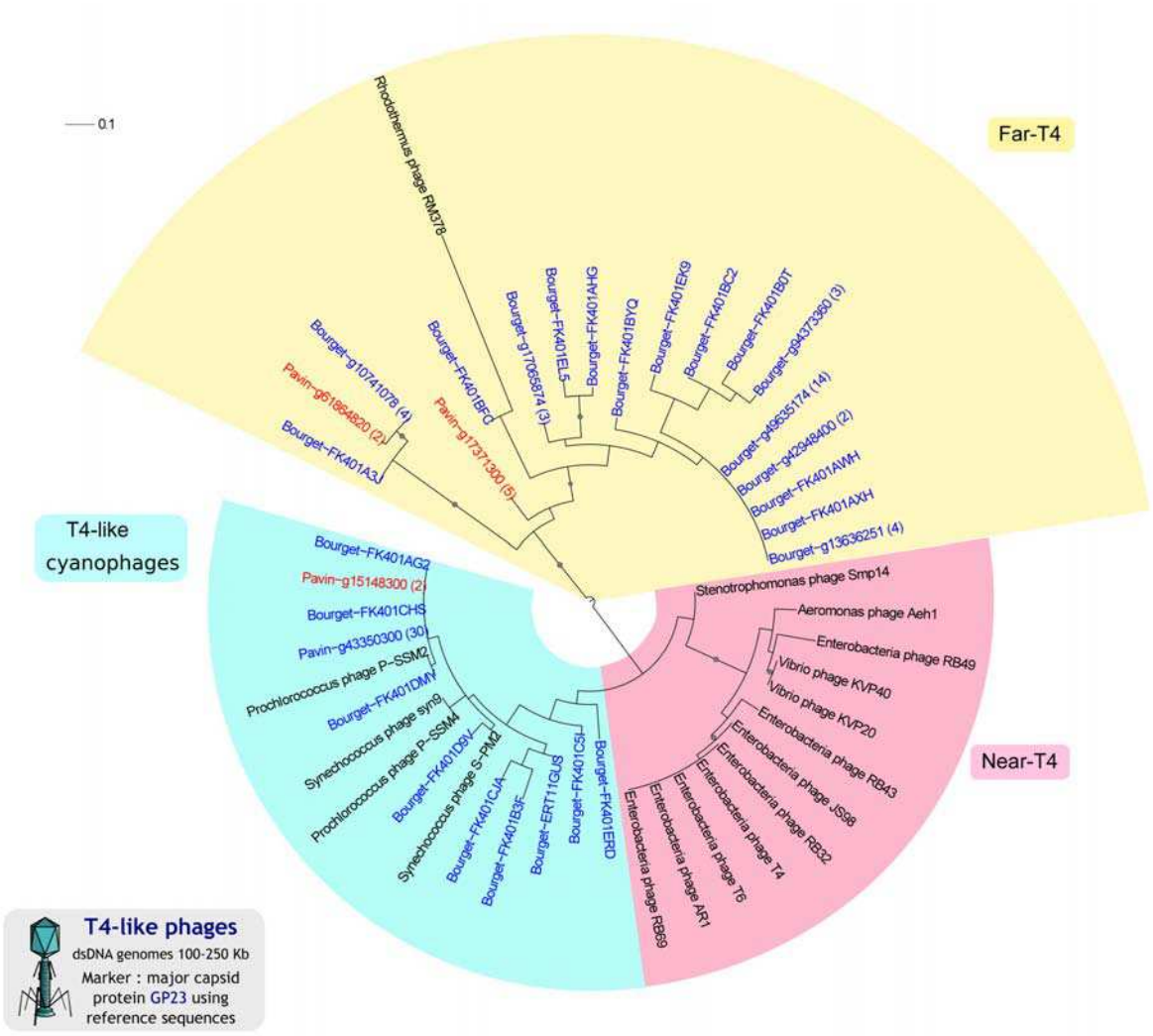
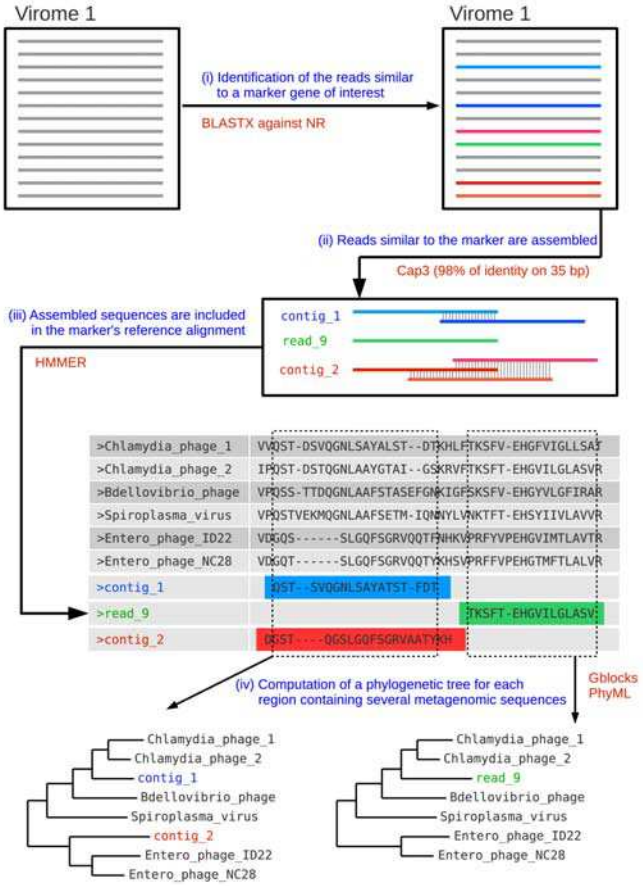


Figure S3 . Maximum-likelihood tree for T4-like phage (G20). A phylogenetic tree has been drawn for the T4-like phage group, and the two main reference groups are indicated (near-T4 in red and T4-like cyanophages in blue). The Far-T4 group is highlighted in yellow. Leaf labels are colored according to their sample (red for Lake Pavin and blue for Lake Bourget). Nodes with at least 80% bootstrap support are flagged with black circles. The sample origin of PCR-obtained sequences are designated on the leaf label (seaw stands for seawater, flood for floodwater, and fresh for freshwater). Rhodothermus RM378, the only cultured representative within the Far-T4 clade, is marked with a black dot.



FigureS4.Maximum-likelihoodtreeforT4-likephages(GP23).
The main reference groups are indicated on the tree (near-T4 in red, T4-like cyanophages in blue), and leaf labels corresponding to virome sequences are colored (red for Lake Pavin and blue for Lake Bourget). The Far-T4 group is highlighted in yellow.
The number of reads assembled is given in brackets foreach contig. Nodes with at least 80% bootstrap support are flagged with black circles.



FigureS5. Schematic representation of the phylogenetic tree creation pipeline.

Annexe A.6 : Matériel supplémentaire

Supplementary material : Meta-analysis of metagenomic data shows that halophilic viral pan-genome is consistent across time and space (Article V)

Sample	Sampling location	Salinity	Dataset description	Nb. of raw reads	Nb. of contigs (>500 nt)	Virome global comparison	Contigs map comparison
P2	Saloon (Senegal)	8.00%	this manuscript	44 532 000	25 749	raw reads	contigs
P5	Ngallou (Senegal)	26.00%	this manuscript	46 905 300	10 904	raw reads	contigs
P6	Lake Retba (Senegal)	29.00%	this manuscript	53 825 100	3 076	raw reads	contigs
P7	Ngallou (Senegal)	29.80%	this manuscript	42 769 700	4 711	raw reads	contigs
P8	Ngallou (Senegal)	35.00%	this manuscript	51 795 500	10 629	raw reads	contigs
P9	Ngallou (Senegal)	36.00%	this manuscript	62 616 100	3 651	raw reads	contigs
Saltern low – 111005	San Diego (USA)	6-8%	Rodriguez-Brito *et al.* (2010)	109 836	-	raw reads	
Saltern low – 112805	San Diego (USA)	6-8%	Rodriguez-Brito *et al.* (2010)	62 363	-	raw reads	
Saltern low – 0704	San Diego (USA)	6-8%	Rodriguez-Brito *et al.* (2010)	268 049	-	raw reads	
Saltern medium – 111605	San Diego (USA)	12-14%	Rodriguez-Brito *et al.* (2010)	58 319	-	raw reads	
Saltern medium – 112205	San Diego (USA)	12-14%	Rodriguez-Brito *et al.* (2010)	55 142	-	raw reads	
Saltern high – 111605	San Diego (USA)	27-30%	Rodriguez-Brito *et al.* (2010)	151 180	-	raw reads	
Saltern high – 120705	San Diego (USA)	27-30%	Rodriguez-Brito *et al.* (2010)	46 628	-	raw reads	
LT_2007_A_1	Lake Tyrell (Australia)	31.00%	Emerson *et al.* (2012)		3 139	raw reads	contigs
LT_2007_A_2	Lake Tyrell (Australia)	31.00%	Emerson *et al.* (2012)		16 328	raw reads	contigs
LT_2010_B_1	Lake Tyrell (Australia)	32.00%	Emerson *et al.* (2012)		7 133	raw reads	contigs
LT_2010_B_2	Lake Tyrell (Australia)	36.00%	Emerson *et al.* (2012)		5 681	raw reads	contigs
LT_2010_B_3	Lake Tyrell (Australia)	34.00%	Emerson *et al.* (2012)		6 360	raw reads	contigs
LT_2010_B_4	Lake Tyrell (Australia)	32.00%	Emerson *et al.* (2012)		12 379	raw reads	contigs
LT_2010_A	Lake Tyrell (Australia)	35.00%	Emerson *et al.* (2012)		20 452	raw reads	contigs
LT_2009_B	Lake Tyrell (Australia)	25.00%	Emerson *et al.* (2012)		27 957	raw reads	contigs
Santa Pola Fosmid	San Diego (USA)	37.00%	Garcia-Hereda *et al.* (2012)	-	42		fosmids

Table S1 : Hypersaline virome datasets used in this study

Taxonomy	P2	P5	P6	P7	P8	P9	Lake Tyrrell 2007 A 1	Lake Tyrrell 2007 A 2	Lake Tyrrell 2009 B	Lake Tyrrell 2010 B 1	Lake Tyrrell 2010 B 2	Lake Tyrrell 2010 B 3	Lake Tyrrell 2010 B 4	Lake Tyrrell 2010 A	Santa Pola fosmids
Viruses	2423	975	210	428	963	157	546	3243	4665	1431	1032	1262	2155	3544	42
dsDNA viruses, no RNA stage	2307	902	187	411	909	151	499	3027	4358	1335	939	1161	1989	3294	42
Caudovirales	1933	698	118	201	555	97	406	2436	3219	1015	729	884	1449	2600	39
Myoviridae	513	234	52	91	195	37	130	744	1221	399	270	351	540	804	8
Tevenvirinae	89	32	6	7	12	3	4	47	96	33	13	29	27	51	0
unclassified Myoviridae	313	152	33	75	159	28	114	628	1042	344	242	298	465	648	8
Natrialba phage PhiCh1	9	32	14	42	90	10	88	355	790	238	182	223	329	373	8
Podoviridae	414	83	16	15	56	3	18	168	146	35	29	34	70	236	0
Siphoviridae	923	359	49	91	294	57	253	1450	1758	569	420	486	822	1509	31
unclassified Siphoviridae	802	303	39	86	271	47	234	1320	1604	522	371	439	741	1368	30
Archaeal BJ1 virus	16	45	15	35	117	17	111	565	689	217	153	161	280	513	23
unclassified Caudovirales	83	22	1	4	10	0	5	74	94	12	10	13	17	51	0
Phycodnaviridae	72	41	6	10	16	11	2	41	219	22	17	15	50	47	0
Salterprovirus	4	26	29	19	40	4	8	91	132	32	35	24	65	126	1
His 1 virus	4	22	26	14	30	3	6	73	111	19	23	10	51	104	0
His 2 virus	0	4	3	5	10	1	2	18	21	13	12	14	14	22	1
unclassified dsDNA phages	188	42	1	7	13	1	6	92	50	15	18	20	33	60	0
unclassified dsDNA viruses	41	64	26	165	269	27	75	331	616	228	126	205	353	410	2
unclassified archaeal dsDNA viruses	20	54	18	147	249	21	75	317	579	224	124	201	339	395	2
Haloarcula hispanica icosahedral virus 2	0	2	0	0	0	0	0	1	1	0	0	0	0	0	0
Haloarcula phage SH1	0	2	0	1	2	0	0	0	1	0	0	0	1	3	0
Haloviruses	18	48	17	144	245	20	75	315	575	224	124	201	338	386	2
Halorubrum phage HF2	0	1	1	2	5	1	1	10	9	2	3	1	4	8	0
Halovirus HF1	5	9	0	33	52	6	13	46	72	48	32	30	48	85	0
Halovirus HSTV-2	0	4	6	14	20	2	11	47	111	46	30	43	69	49	0
Halovirus HVTV-1	13	34	10	95	168	11	50	212	383	128	59	127	217	244	2
Pyrococcus abyssi virus 1	2	2	1	2	2	1	0	1	2	0	0	0	0	6	0
ssDNA viruses	9	5	10	6	8	1	14	40	105	21	23	24	45	37	0
unclassified ssDNA viruses	4	3	6	4	3	1	13	34	101	20	22	23	44	31	0
Halogeometricum pleomorphic virus 1	0	2	2	1	1	0	2	7	23	6	3	3	8	8	0
Halorubrum pleomorphic virus 1	0	0	1	2	1	1	3	3	13	3	8	8	9	4	0
Halorubrum pleomorphic virus 2	0	0	0	0	0	0	3	3	2	2	1	2	1	3	0
Halorubrum pleomorphic virus 3	0	1	2	1	1	0	4	21	62	8	10	10	25	15	0
Halorubrum pleomorphic virus 6	0	0	0	0	0	0	1	0	1	1	0	0	1	1	0
unclassified archaeal viruses	8	1	1	0	3	0	0	1	9	0	1	3	1	7	0
Hyperthermophilic Archaeal Virus 2	1	0	0	0	3	0	0	0	5	0	0	2	0	1	0
Thermococcus prieurii virus 1	7	1	1	0	0	0	0	1	4	0	1	1	1	6	0
unclassified phages	83	57	11	8	26	3	26	147	156	46	44	45	71	165	0

Table S2 : Affiliation of contigs > 500 nt to the species level by best BLAST hit against RefseqVirus.

Id contig	Sample	Contig length (nt)	Type	Fosmid group	TerL Tree group
Environmental Halophage eHP-28	Santa Pola	21 889	fosmid	NC	HCG 1
P5_16	Senegal 5	17 331	linear	-	HCG 1
P7_37	Senegal 7	19 183	linear	-	HCG 1
Environmental Halophage eHP-13	Santa Pola	35 126	fosmid	Cluster3	HCG 2
Environmental Halophage eHP-15	Santa Pola	37 310	fosmid	Cluster4	HCG 2
Environmental Halophage eHP-19	Santa Pola	21 190	fosmid	Cluster4	HCG 2
Environmental Halophage eHP-2	Santa Pola	27 204	fosmid	Cluster1	HCG 2
Environmental Halophage eHP-20	Santa Pola	31 983	fosmid	NC	HCG 2
Environmental Halophage eHP-22	Santa Pola	33 770	fosmid	Cluster1	HCG 2
Environmental Halophage eHP-24	Santa Pola	32 681	fosmid	Cluster1	HCG 2
Environmental Halophage eHP-34	Santa Pola	34 179	fosmid	Cluster4	HCG 2
Environmental Halophage eHP-37	Santa Pola	30 300	fosmid	Cluster1	HCG 2
Environmental Halophage eHP-4	Santa Pola	30 520	fosmid	Cluster3	HCG 2
Environmental Halophage eHP-5	Santa Pola	29 473	fosmid	Cluster1	HCG 2
Environmental Halophage eHP-D7	Santa Pola	31 094	fosmid	Cluster1	HCG 2
Environmental Halophage eHP-E5	Santa Pola	32 692	fosmid	Cluster1	HCG 2
P5_124	Senegal P5	10 311	linear	-	HCG 2
LT_2010_A_10	Lake Tyrrell A	15 318	linear	-	HCG 3
LT_2010_A_11	Lake Tyrrell A	13 799	linear	-	HCG 3
LT_2010_B_1_contig00008	Lake Tyrrell B	10 776	linear	-	HCG 3
P8_9	Senegal P8	31 996	linear	-	HCG 3
P9_55	Senegal P9	19 368	linear	-	HCG 3
Environmental Halophage eHP-23	Santa Pola	31 231	fosmid	Cluster2	HCG 4
Environmental Halophage eHP-35	Santa Pola	31 263	fosmid	Cluster2	HCG 4
LT_2010_B_4_0	Lake Tyrrell B	26 538	circular	-	HCG 4
P8_133	Senegal P8	11 733	linear	-	HCG 4
Environmental Halophage eHP-14	Santa Pola	40 155	fosmid	NC	HCG 5
Environmental Halophage eHP-17	Santa Pola	21 771	fosmid	Cluster6	HCG 5
Environmental Halophage eHP-18	Santa Pola	37 283	fosmid	Cluster6	HCG 5
Environmental Halophage eHP-27	Santa Pola	28 087	fosmid	NC	HCG 5
Environmental Halophage eHP-31	Santa Pola	34 512	fosmid	NC	HCG 5
Environmental Halophage eHP-33	Santa Pola	22 115	fosmid	Cluster6	HCG 5
Environmental Halophage eHP-12	Santa Pola	43 644	fosmid	Cluster5	HCG 6
Environmental Halophage eHP-16	Santa Pola	32 707	fosmid	Cluster5	HCG 6
Environmental Halophage eHP-32	Santa Pola	36 484	fosmid	NC	HCG 6
Environmental Halophage eHP-36	Santa Pola	38 142	fosmid	Cluster5	HCG 6
Environmental Halophage eHP-6	Santa Pola	37 376	fosmid	Cluster5	HCG 6
LT_2010_B_1_contig00001	Lake Tyrrell B	24 192	linear	-	HCG 6
P8_169	Senegal P8	10 496	linear	-	HCG 6
P8_28	Senegal P8	19 284	linear	-	HCG 6
P8_3	Senegal P8	51 314	linear	-	HCG 6

Table S3 : Group and characteristics of long caudovirales contigs. Fosmid groups are taken from the specific study of these sequences (Garcia-Heredia *et al.*, 2012). Hypersaline Caudovirales Groups (HCG) are defined in this study from the tree in Fig. 4 and the contig clustering in Fig. S1.

Coverage						
Virome	Contig	Affiliation	Nb reads	Length	Coverage	Mismatch ratio
Senegal P5	P5_16	HCG 1	22 900	17 331	124.42	0.35%
Senegal P5	P5_124	HCG 2	22 777	10 311	215.42	0.43%
Senegal P7	P7_37	HCG 1	4 602	19 183	22.90	0.39%
Senegal P8	P8_3	HCG 6	47 895	51 314	93.03	0.25%
Senegal P8	P8_9	HCG 3	61 144	31 996	185.92	0.37%
Senegal P8	P8_28	HCG 6	3 548	19 284	18.27	0.48%
Senegal P8	P8_133	HCG 4	3 836	11 733	32.25	0.53%
Senegal P8	P8_169	HCG 6	2 324	10 496	21.81	0.28%
Senegal P9	P9_55	HCG 3	142 799	19 368	731.44	0.31%
Lake Tyrrell 2010 B 4	LT_2010_B_4_0	HCG 4	7 558	26 538	28.39	0.33%
Lake Tyrrell 2010 A	LT_2010_A_10	HCG 3	4 868	15 318	30.6	0.38%
Lake Tyrrell 2010 A	LT_2010_A_11	HCG 3	4 186	13 799	28.5	0.53%

Table S4 : Coverage of large HCG contigs.

Salterprovirus His2

Contigs	Type	Size	**YP_529637 His2V_gp07**
Senegal P8 426	linear	6 230	1
Senegal P8 263	linear	8 505	1
Senegal P8 213	linear	9 520	1
Senegal P5 102	linear	10 918	1
Senegal P7 38	linear	19 002	1
Senegal P8 23	linear	21 217	1
Senegal P5 5	linear	21 301	1
Santa Pola fosmid eHP_34	fosmid	31 983	1
Santa Pola fosmid eHP_20	fosmid	34 179	1

Salterprovirus His1

Contig	Type	Size	**YP_529528 His1V_gp16**	*Archaeal plasmid replication gene : YP_138498 Haloarcula_pSCM201p1*	*Archaeal plasmid primase polymerase domain : cd04859*	*Haloarcula hispanica pleomorphic virus 1 Rep protein : YP_003411995 HHPV-1_gp1*
Senegal P8 1019	circular	3 007	1	1		
Senegal P5 1040	circular	3 388	1	1		
Senegal P5 1022	circular	3 424	1	1		
Senegal P7 384	circular	3 620	1	1		
Senegal P7 365	circular	3 762	1	1		
Senegal P6 125	circular	3 820	1		1	
Senegal P6 120	circular	3 914	1		1	
Senegal P7 338	circular	3 960	1		1	
Senegal P6 117	circular	4 015	1		1	
Senegal P8 629	circular	4 503	1			1
Senegal P6 86	circular	5 025	1			1
Senegal P7 251	circular	5 076	1			1
Senegal P8 514	circular	5 287	1			1
Senegal P8 500	circular	5 362	1			1

Contig	Type	Size	**YP_529528 His1V_gp16**	**YP_529530 His1V_gp18**	**YP_529533 His1V_gp21**	**YP_529538 His1V_gp26**
Senegal P8 243	circular	8 824	1			
Senegal P6 24	circular	11 972	1	1	*	1
Senegal P8 63	circular	14 878	1			
Senegal P8 60	circular	15 156	1			

Contig	Type	Size	**YP_529528 His1V_gp16**	**YP_529533 His1V_gp21**	**YP_529543 His1V_gp31**	**YP_529544 His1V_gp32**
Senegal P7 47	linear	16 482				1
Senegal P8 39	linear	17 335		?	1	1
Senegal P5 13	linear	19 003	1			
Senegal P8 14	circular	23 718	1			
Senegal P6 0	linear	38 687	1			
Senegal P5 0	circular	44 814	1	*	1	1

Table S5 : List of hits for His1 et His2 in contigs and fosmids

Read	Sample	Fraction	Number of gp16 hits
NCBI_ENV_NT_239971678	Hydrothermal vent metagenome LCHCB	microbial	1
NCBI_ENV_NT_256604548	Mine drainage	microbial	3
NCBI_ENV_NT_189042121	Mine drainage	microbial	2
NCBI_ENV_NT_189042097	Mine drainage	microbial	13
NCBI_ENV_NT_189042130	Mine drainage	microbial	1
NCBI_ENV_NT_256601767	Mine drainage	microbial	3
NCBI_ENV_NT_256606635	Mine drainage	microbial	6
NCBI_ENV_NT_256601887	Mine drainage	microbial	6
NCBI_ENV_NT_256601538	Mine drainage	microbial	5
NCBI_ENV_NT_189042115	Mine drainage	microbial	1
NCBI_ENV_NT_181904122	Saltern metagenome 40812378	viral	8
NCBI_ENV_NT_181903599	Saltern metagenome 40828008	viral	3
NCBI_ENV_NT_181904441	Saltern metagenome 40830276	viral	2
NCBI_ENV_NT_180001067	Saltern metagenome 41106982	microbial	2
NCBI_ENV_NT_182211160	Saltern metagenome 42061137	viral	2
NCBI_ENV_NT_178115936	Stromatolite	viral	4
NCBI_ENV_NT_129944359	Mangrove on Isabella Island	microbial	42

Table S6 : List of metagenomic sequences similar to at least one His1_gp16-like gene (tBLASTn of all gp16-like genes against NCBI environmental nucleotide database through the CAMERA web-portal).

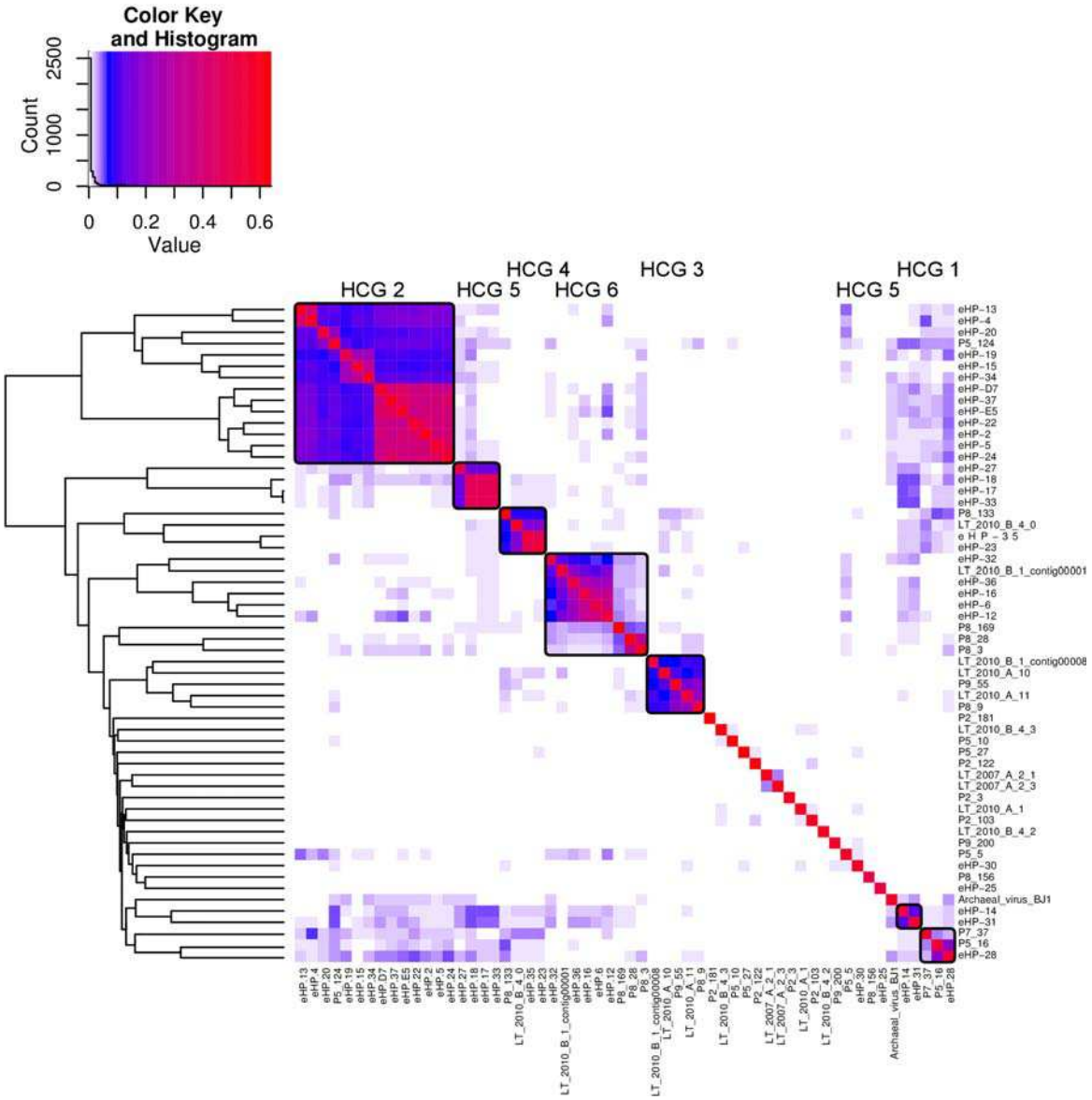


Figure S1 : Clustering of Hypersaline Caudovirales Groups contigs, based on a global similarity matrix computed from BLAST comparison of all predicted ORFs.

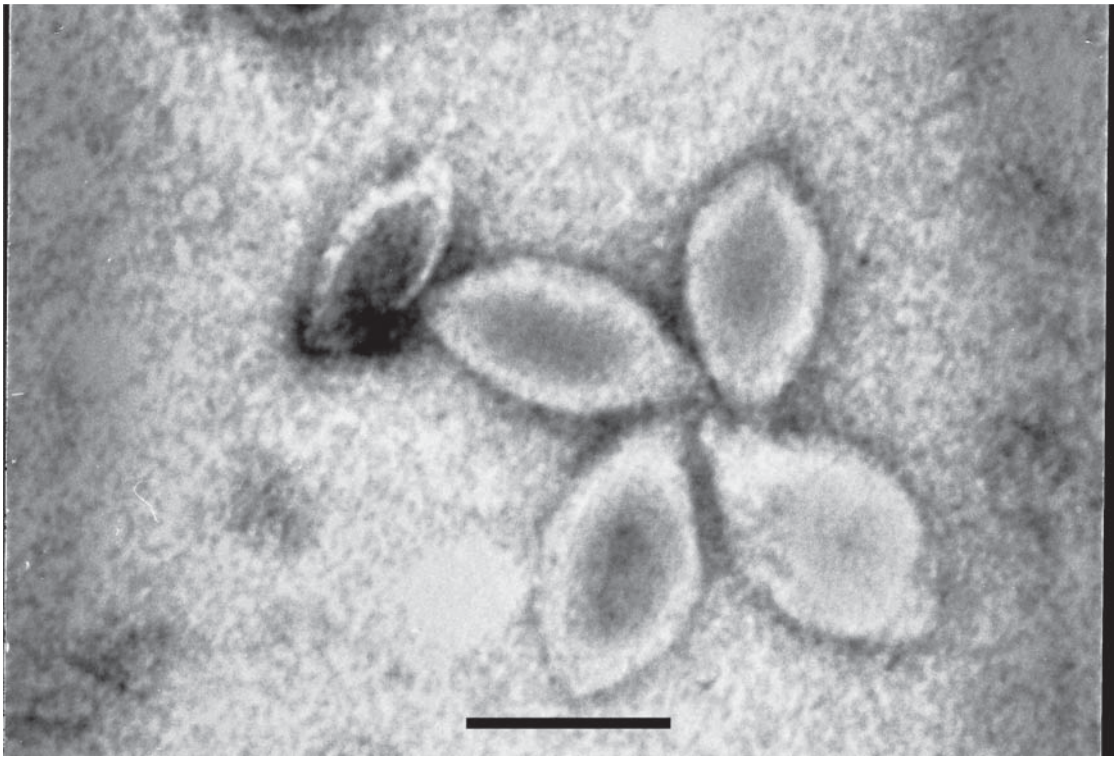


Figure S2 : Transmission electron micrographs of VLPs from sample P7 with lemon-shaped morphologies. Scale bar represents 100 nm.

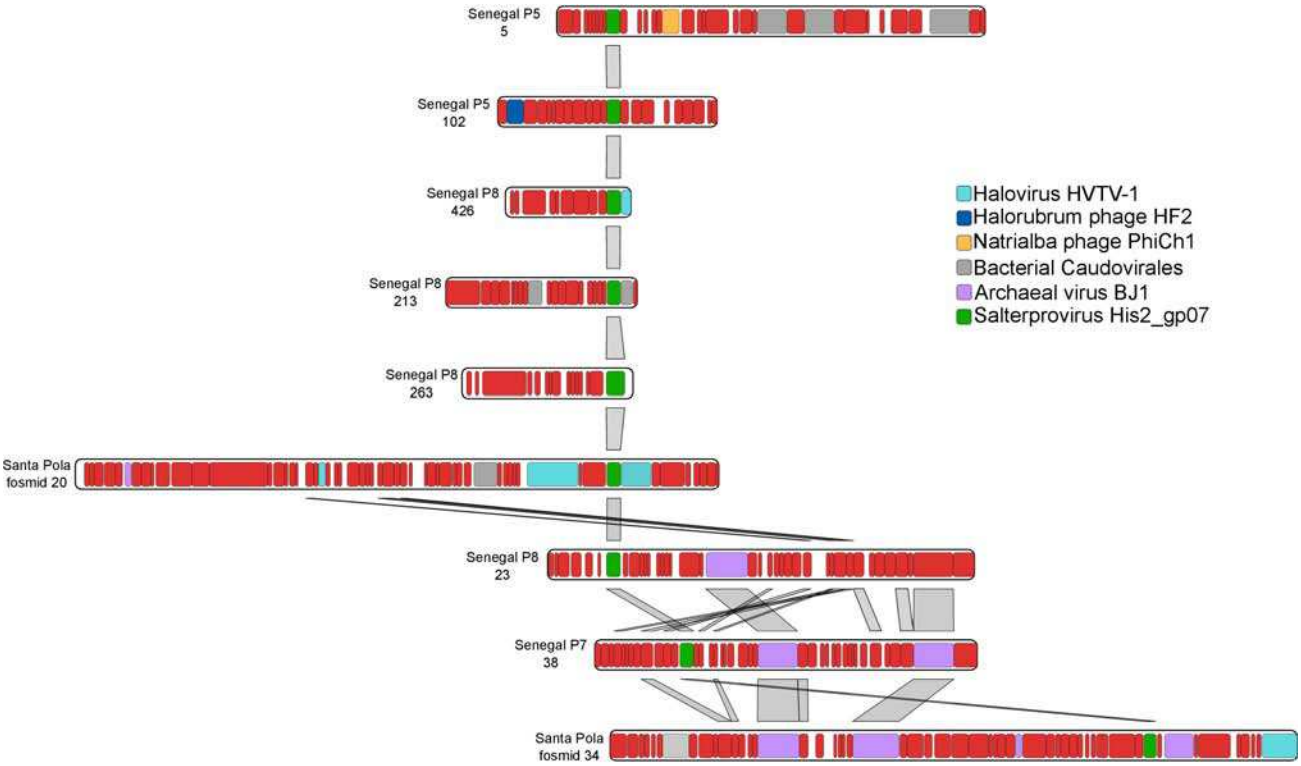


Figure S3 : Map comparison for contigs containing a predicted gene similar to His2 argynil tRNA synthetase (His2-gp07).

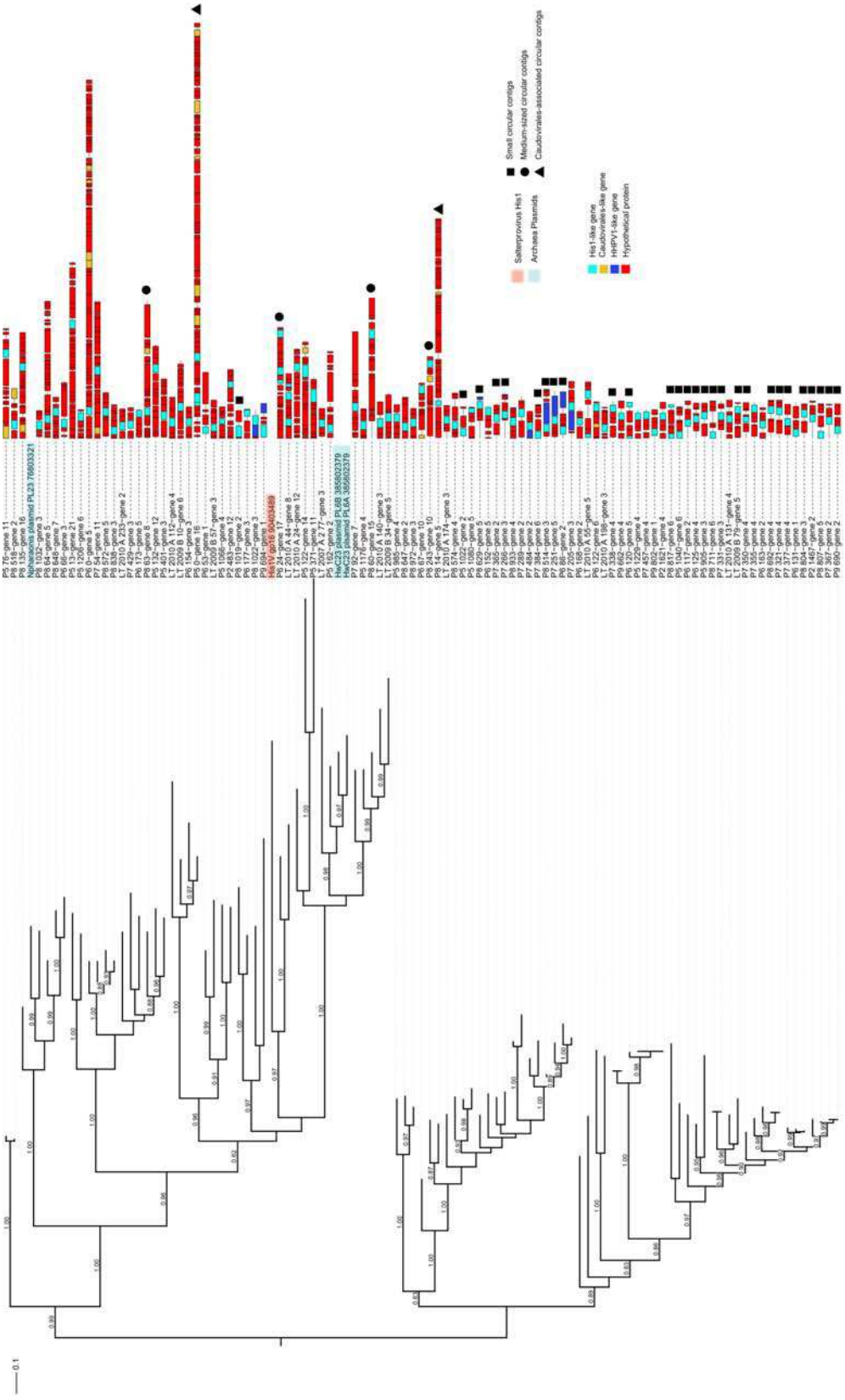


Figure S4 : Phylogenetic tree based on His1-gp16 like sequences. Bootstraps scores are indicated on each node. HHPV1 stands for *Haloarcula Hispanica* Pleomorphic Virus 1.

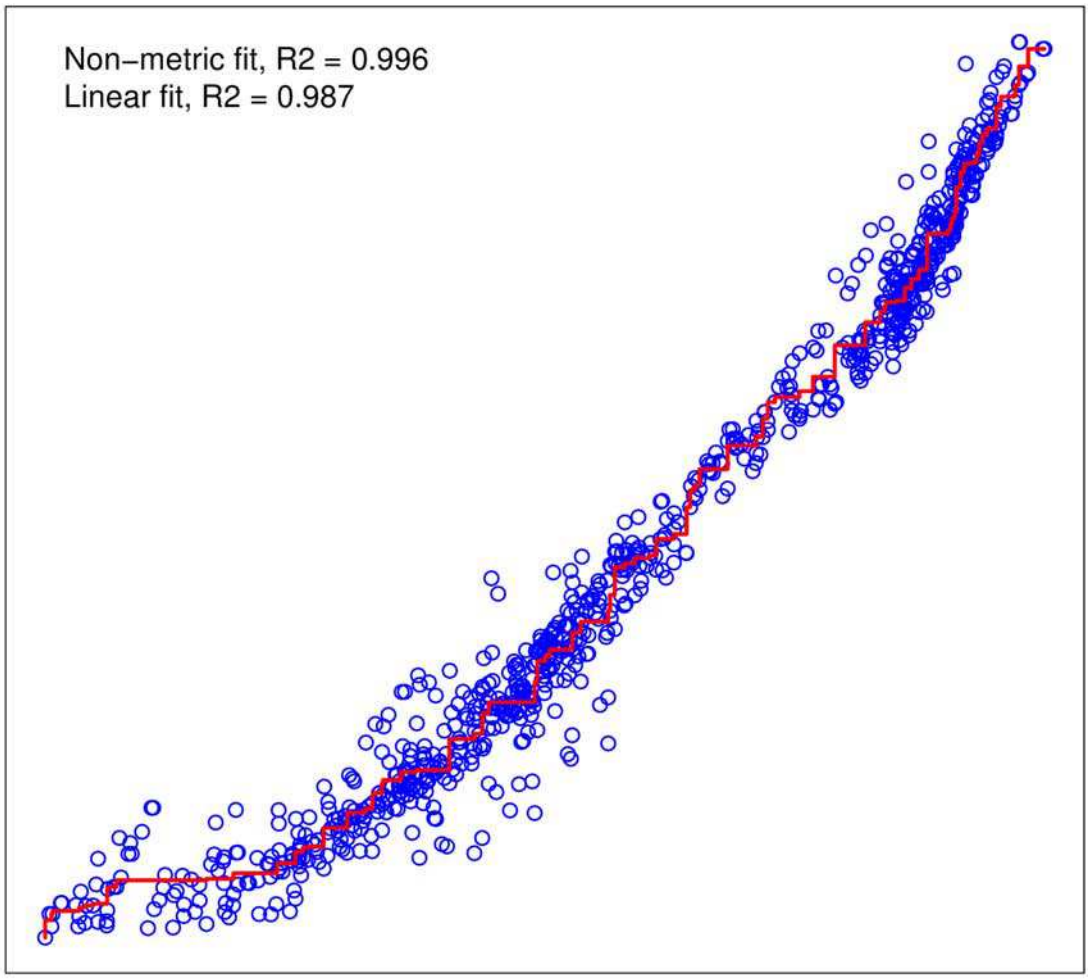


Figure S5 : Shepard plot of virome comparison MDS. For each pairwise comparison, the euclidean distance computed from the BLAST-based score matrix is reported on x-axis, and the distance on the MDS plot on the y-axis.

Annexe A.7 : Matériel supplémentaire

Supplementary material : Evolution and diversity of the *Microviridae* viral family through a collection of 81 new complete genomes assembled from virome reads (Article VI)

Virome	Available on	Methodology used in sample preparation	Sample origin	Sample type	Number of sequences	Mean size of sequences	Contigs with VP1 hits detected as circular
Lake Pavin	Metavir – Project French Lakes	PEG – MDA	Lake Pavin – France	Freshwater	649,290	412.31	3
Lake Bourget	Metavir – Project French Lakes	PEG – MDA	Lake Bourget – France	Freshwater	593,084	433.41	12
12 Saline MedSalternSDbayVir112805	MG-Rast – 4440427.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	39,439	100.43	
13 Saline MedSalternSDbayVir111605	MG-Rast – 4440428.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	58,319	98.11	
14 Saline HighSalternSDbayVir111605	MG-Rast – 4440421.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	151,180	99.76	
15 Saline Low SalternSDbayVir0704	MG-Rast – 4440436.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	268,049	104.47	
16 Saline Low SalternSDbayVir111005	MG-Rast – 4440432.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	109,836	104.35	
17 Saline MedSalternSDbayVir111005	MG-Rast – 4440431.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	39,348	101.40	
18 Saline MedSalternSDbayVir112205	MG-Rast – 4440417.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	55,142	100.79	
19 Saline HighSalternSDbayVir120705	MG-Rast – 4440145.4	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	46,628	102.15	
20 Saline HighSalternSDbayVir112805	MG-Rast – 4440144.4	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	4,536	100.15	
21 Saline Low SalternSDbayVir112805	MG-Rast – 4440420.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	62,363	103.87	
22 Saline SaltonSeaVirOne082308	MG-Rast – 4440327.3		Salton Sea – California	Hypersaline	55,467	103.60	
23 Saline SaltonSeaVirTw o082308	MG-Rast – 4440328.3		Salton Sea – California	Hypersaline	29,814	99.20	
32 Marine GOMVir94to01	MG-Rast – 4440304.3	CsCL – MDA	Gulf of Mexico	Seawater	262,501	101.59	
33 Marine BBCVir96to04	MG-Rast – 4440305.3	CsCL – MDA	British Columbia	Seawater	414,964	102.14	
34 Marine ArcticVir2002	MG-Rast – 4440306.3	CsCL – MDA	Arctic sea	Seawater	686,209	99.15	
35 Marine SARVir063005	MG-Rast – 4440322.3	CsCL – MDA	Sargasso sea	Seawater	397,939	104.31	*
36 CoralAtol KingLIVir082105	MG-Rast – 4440036.3	CsCL – MDA	Kingmann - Line Islands	Seawater	93,744	108.34	
37 CoralAtol XmasLIVir080505	MG-Rast – 4440038.3	CsCL – MDA	Christmas - Line Islands	Seawater	279,882	110.56	
38 CoralAtol PalmLIVir081805	MG-Rast – 4440040.3	CsCL – MDA	Palmyra - Line Islands	Seawater	318,178	104.78	
39 CoralAtol FannLIVir081105	MG-Rast – 4440280.3	CsCL – MDA	Tabuaeran - Line Islands	Seawater	378,475	104.13	
40 MarineBay 4440102.3	MG-Rast – 4440102.3	PEG – CsCL- MDA	Tampa Bay – Florida	Seawater	279,129	103.95	
41 MarineBay SkanBayAKVir092706	MG-Rast – 4440330.3		Skan Bay – Alaska	Seawater	30,831	104.56	
46 Fresh TilPondKentSTVir0806	MG-Rast – 4440424.3	PEG – CsCL- MDA	Tilapia Pond 3 – California	Freshwater	56,549	101.06	
47 Fresh TilPondKentSTVir050406	MG-Rast – 4440412.3	PEG – CsCL- MDA	Healthy fish Pond – California	Freshwater	60,135	101.07	
48 Fresh PrePondKentSTVir050406	MG-Rast – 4440414.3	PEG – CsCL- MDA	Prebead Pond – California	Freshwater	67,785	103.21	
49 Fresh TpondKentSTVir1105	MG-Rast – 4440439.3	PEG – CsCL- MDA	Tilapia Pond – California	Freshwater	264,844	102.25	
57 Coral T0PortComHaw Vir022306	MG-Rast – 4440376.3	CsCL – MDA	Porites Compressa	Eukaryote	39,113	101.32	
58 Coral ConPorCompHaw Vir0206	MG-Rast – 4440374.3	CsCL – MDA	Porites Compressa	Eukaryote	39,191	103.70	
59 Coral DOCPorCompVirHaw 0206	MG-Rast – 4440370.3	CsCL – MDA	Porites Compressa	Eukaryote	35,409	102.18	1
60 Coral pHPorCompHaw Vir0206	MG-Rast – 4440371.3	CsCL – MDA	Porites Compressa	Eukaryote	49,949	104.73	
61 Coral NutPorCompHaw Vir0206	MG-Rast – 4440377.3	CsCL – MDA	Porites Compressa	Eukaryote	34,139	107.18	
62 Coral TempPorCompHaw Vir0206	MG-Rast – 4440375.3	CsCL – MDA	Porites Compressa	Eukaryote	38,482	113.38	
66 Microbialites PAStromCCMexVir072205	MG-Rast – 4440320.3	CsCL – MDA	Paztac Azules	Microbialites	301,264	104.64	1
67 Microbialites RMStromCCMexVir072205	MG-Rast – 4440321.3	CsCL – MDA	Rio Mesquites	Microbialites	324,500	104.22	
68 Microbialites HBCStromBahamasVir011105	MG-Rast – 4440323.3	CsCL – MDA	Bahamas	Microbialites	148,334	100.52	1
73 Fish FishHealSlimKentSTVir050406	MG-Rast – 4440065.3		Healthy fish gut	Eukaryote	61,022	98.45	
74 Fish FishMorSlimKentSTVir050406	MG-Rast – 4440064.3		Morbid fish gut	Eukaryote	59,599	98.32	
84 Animal HealSputSDRep3Vir070706	MG-Rast – 4440442.4	CsCL – MDA	Mosquito (USA)	Eukaryote	39,489	84.80	
85 Mosquito MosqISDVir01252006	MG-Rast – 4440052.3	CsCL – MDA	Mosquito (USA)	Eukaryote	336,760	102.61	
86 Mosquito MosqDigSDVir060606	MG-Rast – 4440053.3	CsCL – MDA	Mosquito (USA)	Eukaryote	638,689	100.32	
87 Mosquito MosqIISDVir060606	MG-Rast – 4440054.3	CsCL – MDA	Mosquito (USA)	Eukaryote	601,040	104.16	
Antarctic Lake LopezBueno Spring	Metavir – Project Lake Limnopolar	Sucrose cushion – MDA	Lake Limnopolar	Freshwater	41,322	239.65	
Antarctic Lake LopezBueno Summer	Metavir – Project Lake Limnopolar	Sucrose cushion – MDA	Lake Limnopolar	Freshwater	38,475	221.27	
Bear Paw JGI AOIX	MG-Rast – 4441095.3	LASL	Yellowstone Park	Hot Spring	8,352	990.28	
CF10LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	188,287	217.42	
CF10MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	31,411	207.33	
CF6LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	156,809	217.66	
CF6MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	29,071	203.21	
CF7LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	184,666	228.15	
CF7MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	23,689	212.50	1
CF8LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	158,912	246.84	
CF8MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	57,780	216.48	
CF9LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	233,854	205.68	
CF9MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	32,798	194.92	
JCVI SMPL 1103283000058 mv858	Metavir – Virome GOS Move858	.22 µm filtration	Chesapeake Bay – Maryland	Seawater	11,496	1014.07	2
Norm3LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	245,109	220.94	
Norm3MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	47,250	210.54	
Norm4LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	235,755	212.00	
Norm4MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	14,514	206.69	
Norm5LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	211,401	226.10	
Norm5MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	52,680	216.18	
Norm6LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	293,359	219.66	
Norm6MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	141,415	204.02	
Norm7LungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	284,384	210.21	
Norm7MLungVir20080407	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	34,764	204.43	
Octopus JGI	MG-Rast – 4442583.3	LASL	Yellowstone Park	Hot Spring	22,272	1012.80	
SectLung2LLL-PVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	8,981	391.39	1
SectLung2LMLVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	14,037	351.82	
SectLung2LULVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	14,559	398.31	
SectLung2RLLVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	9,232	387.51	
SectLung2RMLVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	5,882	304.36	
SectLung2RULVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	9,026	368.91	
Human Gut 21	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – X1	Eukaryote	30,873	372.90	2
Human Gut 22	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L1	Eukaryote	132,569	379.12	1
Human Gut 23	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	61,104	370.82	
Human Gut 24	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	148,781	370.31	1
Human Gut 25	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	16,955	375.02	
Human Gut 26	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	13,048	378.90	
Human Gut 27	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	16,747	365.41	2
Human Gut 28	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H2	Eukaryote	16,137	366.71	
Human Gut 30	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L1	Eukaryote	107,259	405.06	2
Human Gut 31	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	107,993	409.48	4
Human Gut 32	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	25,648	377.67	3
Human Gut 33	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H2	Eukaryote	23,614	372.91	5
Human Gut 34	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	33,489	369.66	1
Human Gut 35	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	76,090	384.09	1
Human Gut 36	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	15,166	382.16	1
Human Gut 37	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	59,155	382.72	1
Human Feces A	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	113,054	431.05	11
Human Feces B	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	109,569	435.21	7
Human Feces C	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	68,391	437.07	6
Human Feces D	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	115,121	433.08	6
Human Feces E	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	98,511	417.05	5
Total			-	-	12,731,456	195	81

Table S1 : List of viromes assembled. For each virome, the number of circular contigs identified as complete *Microviridae* genome is indicated. The web-server hosting the datasets are : NCBI (www.ncbi.nlm.nih.gov) MG-Rast (http://metagenomics.anl.gov), and Metavir (http://metavir-meb.univ-bpclermont.fr). When available, the methodology used to purify viral particle is indicated (CsCl : Cesium Chloride, PEG : Polyethylene Glycol, LASL : linker amplified shotgun library and MDA : phi29-mediated multiple displacement amplification).

*2 contigs were detected for virome 35 Marine_Sar_Vir, but they corresponded to the 2 contigs already assembled from this virome, described in Tucker et al., 2011.

Contig	Clade	Size (bp)	Coverage	Number of reads assembled	Major capsid protein (VP1) best BLAST hit (gi / bit_score)	Replication protein (VP4) best BLAST hit (gi / bit_score)	DNA Pilot protein (VP2) best BLAST hit (gi / bit_score)	Number of other ORFs predicted
Pavin_279	*Pichovirinae*	3,992	20.13	209	12085136 / 291	9791179 / 99.8	198274650 / 48.5	2
Pavin_723	*Pichovirinae*	3,989	46.52	474	313766927 / 309	9791179/ 104	325269158 / 47	4
Pavin_110	*Gokushovirinae*	4,194	17.37	178	313766927 / 547	313766930 / 118	47522479 / 89.7	3
Bourget_523	*Pichovirinae*	4,193	237.14	2471	313766927 / 323	137995 / 103	17402852 / 33.1	1
Bourget_052	*Gokushovirinae*	4,296	26.03	274	313766927 / 600	313766930 / 106	77020116 / 61.2	4
Bourget_154	*Gokushovirinae*	4,219	249.34	2607	313766927 / 566	313766930 / 121	7406592 / 87.8	2
Bourget_164	*Gokushovirinae*	4,228	125.83	1298	313766927 / 565	313766924 / 131	17402852 / 92	2
Bourget_224	*Gokushovirinae*	4,200	83.44	871	313766927 / 560	313766930 / 169	9791177 / 48.1	3
Bourget_245	*Gokushovirinae*	4,165	28.47	299	313766927 / 564	313766930 / 152	9791177 / 72.4	2
Bourget_248	*Gokushovirinae*	4,281	198.23	2160	313766927 / 571	313766930 / 256	7406592 / 63.5	0
Bourget_259	*Gokushovirinae*	4,149	7.82	82	7190965 / 574	313766930 / 153	9791177 / 62.8	2
Bourget_309	*Gokushovirinae*	4,197	10.93	116	313766927 / 566	313766924 / 108	77020116 / 71.2	4
Bourget_332	*Gokushovirinae*	4,289	11.00	122	313766927 / 558	313766924 / 129	47522479 / 88.6	3
Bourget_504	*Gokushovirinae*	4,209	279.52	2852	313766927 / 545	313766930 / 254	77020116 / 66.6	0
Bourget_915	*Gokushovirinae*	4,285	150.13	1631	313766927 / 557	313766924 / 132	7406592 / 85.1	3
59_Coral_002	*Pichovirinae*	4,195	36.15	1426	313766927 / 474	137995 / 202	218131118 / 67	2
66_Microbialite_001	*Pichovirinae*	4,436	57.65	2488	313766927 / 571	137995 / 253	218131118 / 117	0
68_Microbialite_003	*Gokushovirinae*	4,660	34.32	1611	12085136 / 700	75089164 / 365	75089169 / 44.7	1
CF7ML_001	*Alpavirinae*	6,723	115.88	3669	217762 / 61.6	137995 / 52.4	218131118 / 41.6	0
JCVI_001	*Gokushovirinae*	4,125	19.07	87	313766927 / 805	313766930 / 345	47566144 / 92.8	0
JCVI_003	*Pichovirinae*	4,240	9.13	47	313766927 / 560	137995 / 197	218131118 / 71.6	2
SectLung2LLL_002	*Gokushovirinae*	4,196	11.94	125	313766927 / 638	313766930 / 245	47566144 / 60.1	2
Human_gut_21_005	*Alpavirinae*	6,399	51.70	877	75089173 / 124	7406595 / 90.1	198274650 / 47.8	3
Human_gut_21_019	*Gokushovirinae*	5,029	64.16	911	313766927 / 520	313766930 / 233	9629153 / 34.3	3
Human_gut_22_017	*Alpavirinae*	6,251	11.10	200	313766927 / 82	137995 / 61.6	198274650 / 40	3
Human_gut_24_085	*Alpavirinae*	5,281	31.78	470	313766927 / 62	137995 / 61.2	-	3
Human_gut_27_015	*Gokushovirinae*	5,035	39.86	587	313766927 / 503	77020116 / 45.4	313766930 / 229	4
Human_gut_27_035	*Gokushovirinae*	5,266	21.31	335	313766927 / 458	313766930 / 187	75089169 / 56.6	3
Human_gut_30_017	*Alpavirinae*	5,755	27.81	428	313766927 / 57	313766930 / 60.1	198274650 / 34.7	1
Human_gut_30_040	*Alpavirinae*	6,061	31.71	500	313766927 / 57	9791179 / 89.7	237718677 / 37.4	3
Human_gut_31_037	*Alpavirinae*	6,166	11.79	194	75089173 / 79.7	137995 / 72	218131118 / 45.8	7
Human_gut_31_045	*Gokushovirinae*	5,266	69.40	963	313766927 / 458	313766930 / 187	75089169 / 56.6	3
Human_gut_31_054	*Gokushovirinae*	5,036	36.20	478	313766927 / 503	313766930 / 229	77020116 / 43.5	3
Human_gut_31_126	*Alpavirinae*	5,696	32.27	454	313766927 / 82	137995 / 46.2	218131118 / 38.3	1
Human_gut_32_012	*Alpavirinae*	5,695	90.70	1467	313766927 / 82	137995 / 46.2	218131118 / 39.3	1
Human_gut_32_015	*Alpavirinae*	5,279	41.75	595	313766927 / 59.7	9791179 / 52.8	-	3
Human_gut_32_030	*Gokushovirinae*	5,266	22.63	341	313766927 / 458	313766930 / 187	75089169 / 56.6	3
Human_gut_33_003	*Gokushovirinae*	5,704	94.80	1549	313766927 / 389	313766924 / 138	77020116 / 61.2	3
Human_gut_33_005	*Alpavirinae*	6,449	12.23	222	313766927 / 88.2	137995 / 62.4	218131118 / 35.4	1
Human_gut_33_017	*Alpavirinae*	6,171	138.55	2424	313766927 / 75.9	137995 / 62	9634952 / 38.1	1
Human_gut_33_018	*Gokushovirinae*	4,799	8.99	115	9634949 / 529	313766930 / 200	77020116 / 36.2	1
Human_gut_33_023	*Gokushovirinae*	4,549	23.97	302	313766927 / 604	313766930 / 238	9634952 / 38.5	1
Human_gut_34_012	*Gokushovirinae*	4,822	35.70	484	47566141 / 566	313766930 / 215	313766929 / 39.3	1
Human_gut_35_025	*Gokushovirinae*	4,822	371.64	4986	47566141 / 566	313766930 / 215	313766929 / 39.3	1
Human_gut_36_019	*Gokushovirinae*	4,822	58.45	788	47566141 / 567	313766930 / 215	313766929 / 39.3	1
Human_gut_37_015	*Gokushovirinae*	4,822	84.69	1179	47566141 / 566	313766930 / 215	313766929 / 39.3	1
Human_feces_A_013	*Gokushovirinae*	5316	14.07	177	77020115 / 434	77020121 / 137	12085140 / 53.1	1
Human_feces_A_016	*Alpavirinae*	5914	74.86	1033	288927736 / 43.9	137995 / 53.1	9634952 / 31.2	4
Human_feces_A_019	*Gokushovirinae*	5663	206.03	2881	9634949 / 389	77020121 / 138	12085140 / 46.6	3
Human_feces_A_020	*Gokushovirinae*	5776	32.76	463	9634949 / 375	77020121 / 136	12085140 / 46.6	2
Human_feces_A_021	*Alpavirinae*	6310	38.10	579	237722024 / 399	237722023 / 92	198274650 / 66.2	2
Human_feces_A_029	*Gokushovirinae*	5158	10.10	119	77020115 / 381	77020121 / 143	12085140 / 65.5	3
Human_feces_A_032	*Alpavirinae*	5818	1023.43	14986	270340022 / 59.3	47566147 / 59.3	218131118 / 50.4	3
Human_feces_A_033	*Alpavirinae*	5502	733.10	9647	270340022 / 49.7	137995 / 31.6	218131118 / 30.4	2
Human_feces_A_034	*Alpavirinae*	6301	80.14	1245	198274652 / 332	237722023 / 135	237722026 / 62.4	3
Human_feces_A_047	*Alpavirinae*	5893	33.93	503	270340022 / 53.9	47566147 / 53.1	218131118 / 37	5
Human_feces_A_048	*Alpavirinae*	5816	85.90	1209	154795174 / 52.8	19424731 / 48.1	218131114 / 47	5
Human_feces_B_007	*Alpavirinae*	6233	360.34	518	77020115 / 100	269303098 / 82.8	282877222 62.4	0
Human_feces_B_020	*Alpavirinae*	6041	153.95	2264	270340022 / 57.4	313766930 / 61.2	71843186 / 29.6	6
Human_feces_B_021	*Alpavirinae*	5910	104.89	1510	154795172 / 51.6	9634955 / 50.1	218131114 / 47.4	5
Human_feces_B_023	*Alpavirinae*	5733	118.99	1635	270340022 / 66.6	137995 / 44.7	47522479 / 30.0	4
Human_feces_B_029	*Gokushovirinae*	5226	62.51	771	17402851 / 492	9791179 / 121	77020116 / 61.6	0
Human_feces_B_039	*Alpavirinae*	6008	280.01	4085	237722024 / 293	237722023 / 104	9634952 / 30.8	0
Human_feces_B_068	*Gokushovirinae*	5123	10.29	142	77020115 / 446	19424731 / 119	12085140 / 59.3	1
Human_feces_C_010	*Alpavirinae*	6364	38.05	559	9634949 / 95.1	9634955 / 90.5	198274650 / 43.1	2
Human_feces_C_014	*Gokushovirinae*	5536	17.46	232	77020115 / 357	9791179 / 123	212709260 / 137	1
Human_feces_C_016	*Alpavirinae*	5985	89.92	1222	270340022 / 45.4	9791179 / 75.5	47566147 / 48.1	1
Human_feces_C_029	*Alpavirinae*	5888	32.48	452	270340022 / 62.1	313766932 / 51.2	218131118 / 46.2	4
Human_feces_C_031	*Gokushovirinae*	4810	57.94	668	17402851 / 602	47566147 / 225	12085142 / 73.9	1
Human_feces_C_043	*Alpavirinae*	6226	57.59	863	218131116 / 412	237722023 / 117	218131118 / 49.3	2
Human_feces_D_008	*Alpavirinae*	6582	15.30	253	218131116 / 390	237722023 / 157	218131118 / 69.3	1
Human_feces_D_014	*Gokushovirinae*	5274	109.19	1427	47566141 / 479	77020121 / 165	9791177 / 49.7	1
Human_feces_D_015	*Alpavirinae*	5231	43.03	639	218131116 / 407	218131115 / 88.6	218131118 / 61.2	0
Human_feces_D_022	*Alpavirinae*	5481	1038.68	14358	270340022 / 53.1	137995 / 29.6	325269158 / 44.3	1
Human_feces_D_031	*Alpavirinae*	6659	239.97	3967	218131116 / 387	237722023 / 158	218131118 / 60.8	3
Human_feces_D_045	*Gokushovirinae*	5299	36.11	468	9791178 / 322	19424731 / 125	12085142 / 42.7	1
Human_feces_E_007	*Gokushovirinae*	5774	124.23	1862	9791178 / 351	9634955 / 122	9791177 / 59.7	4
Human_feces_E_009	*Gokushovirinae*	4928	591.76	7568	47566141 / 453	47566147 / 160	12085140 / 57.8	0
Human_feces_E_010	*Gokushovirinae*	4881	425.09	5561	237722024 / 390	237722023 / 89.4	218131118 / 60.8	0
Human_feces_E_011	*Alpavirinae*	6313	61.09	1006	17402851 / 378	77020121 / 140	12085140 / 46.6	3
Human_feces_E_017	*Gokushovirinae*	5769	50.22	755	77020115 / 605	47566147 / 187	9791177 / 75.5	2

Table S2 : List of circular contigs similar to complete genomes of *Microviridae.* For each major protein, the gi of the best BLAST hit is indicated with the bit score of the corresponding BLAST. All the sequences and corresponding annotations are available through Dryad Digital Repository, accession number doi:10.5061/dryad.8ht80

Genome		NR best BLAST hit			
Sug-group	Contig	Gene name	NCBI accession number	Score	E-value
Alpavirinae	CF7ML001	hypothetical protein CFU_3358 [Collimonas fungivorans Ter331]	YP_004754005	100	4E-25
Alpavirinae	Human_gut_22_017	hypothetical protein CFU_3358 [Collimonas fungivorans Ter331]	YP_004754005	52.8	2E-07
Alpavirinae	Human_gut_33_005	peptidase M15A [Leptothrix cholodnii SP-6]	YP_001793077	103	1E-25
Alpavirinae	Human_gut_33_017	hypothetical protein BDI_0132 [Parabacteroides distasonis ATCC 8503]	YP_001301549	52.4	3E-07
Alpavirinae	Human_feces_A_016	putative peptidase [Bacteroides phage B40-8]	YP_002221548	87.4	8E-21
Alpavirinae	Human_feces_B_039	peptidase M15 family protein [Prevotella veroralis F0319]	ZP_05856140	56.6	1E-08
Gokushovirinae	Human_feces_C_014	hypothetical protein PROVALCAL_00295 [Providencia alcalifaciens DSM 30120]	ZP_03317388	137	2E-40
Gokushovirinae	Human_gut_33_003	hypothetical protein FAEPRAM212_00204 [Faecalibacterium prausnitzii M21/2]	ZP_02089970	122	2E-34
Gokushovirinae	Human_feces_A_020	peptidase M15 [Faecalibacterium cf. prausnitzii KLE1255]	ZP_07798563	97.8	1E-24
Gokushovirinae	Human_feces_E_007	peptidase M15 [Faecalibacterium cf. prausnitzii KLE1255]	ZP_07798563	103	7E-27
Gokushovirinae	Human_feces_E_017	peptidase M15 [Faecalibacterium cf. prausnitzii KLE1255]	ZP_07798563	108	9E-29

Table S3. List of the *Microviridae* peptidase genes detected, with their best BLAST hit against NR database.

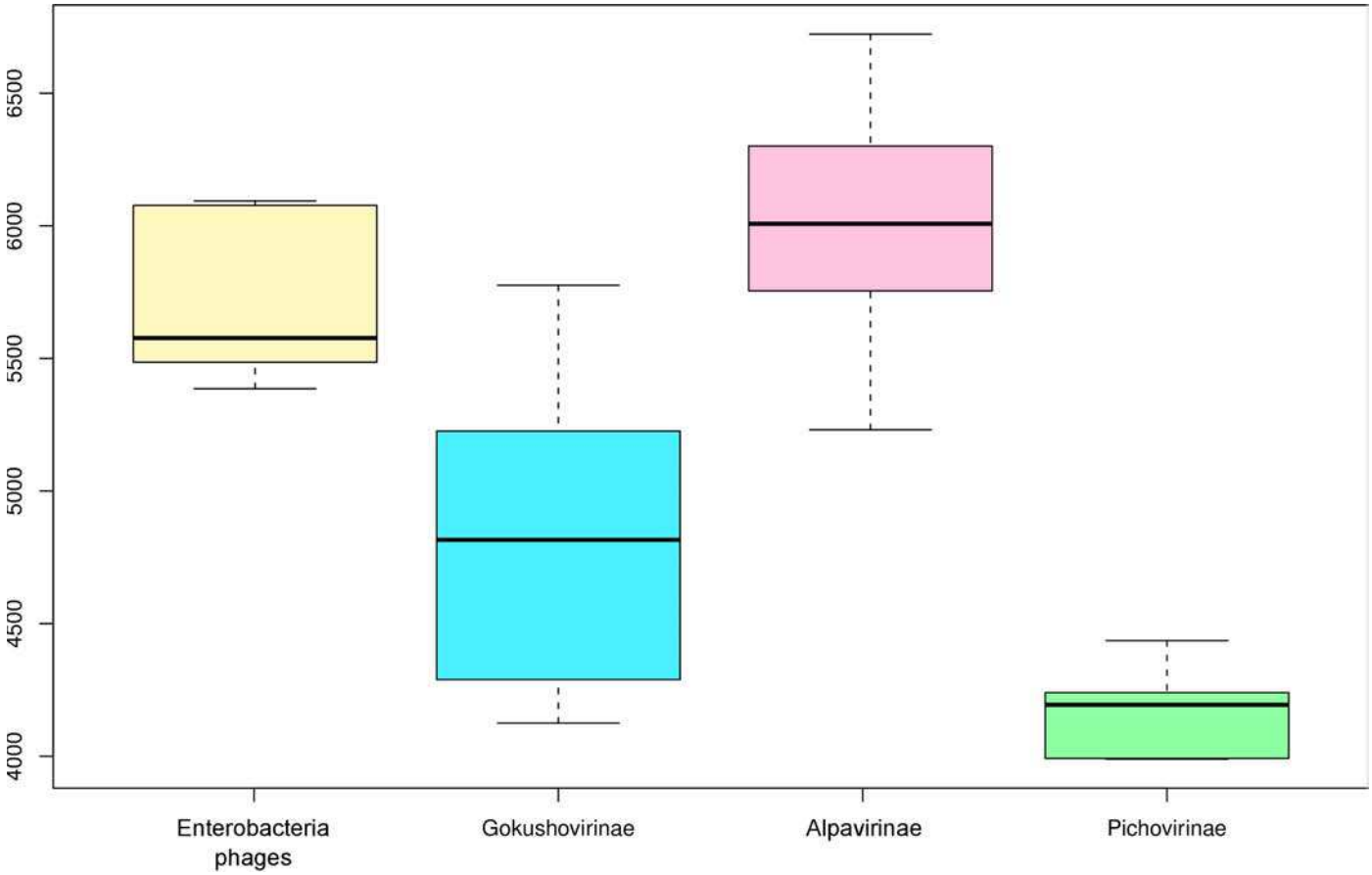


Figure S1. Boxplot of genome sizes within each clade. Affiliations were based on the major capsid protein phylogenetic tree (Fig 1).

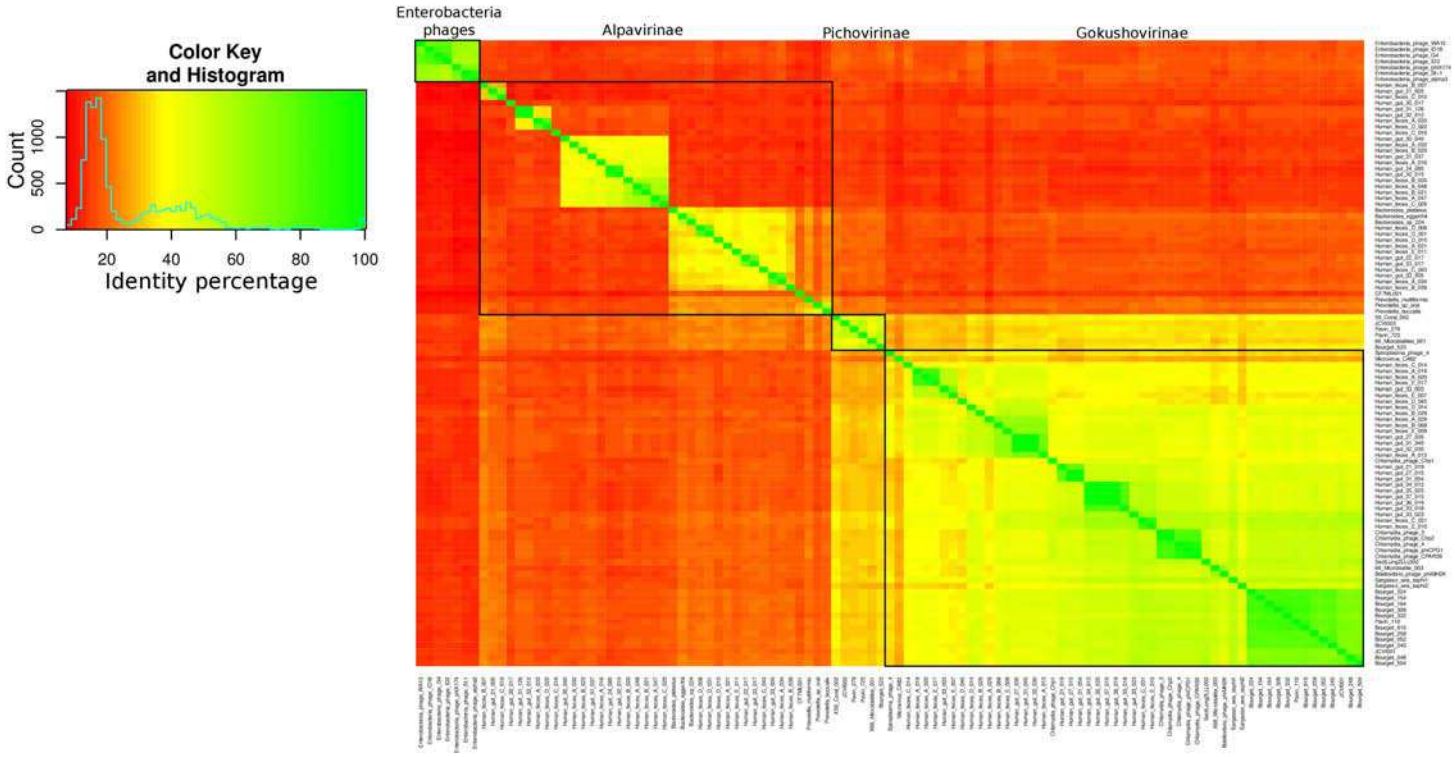


Figure S2 : Heatmap based on the percentage of identity computed from the major capsid protein multiple alignment. Scale is indicated on the top left, with the distribution of the percentages of identity. The genome affiliation is indicated above the map, and groups are framed on the heatmap.

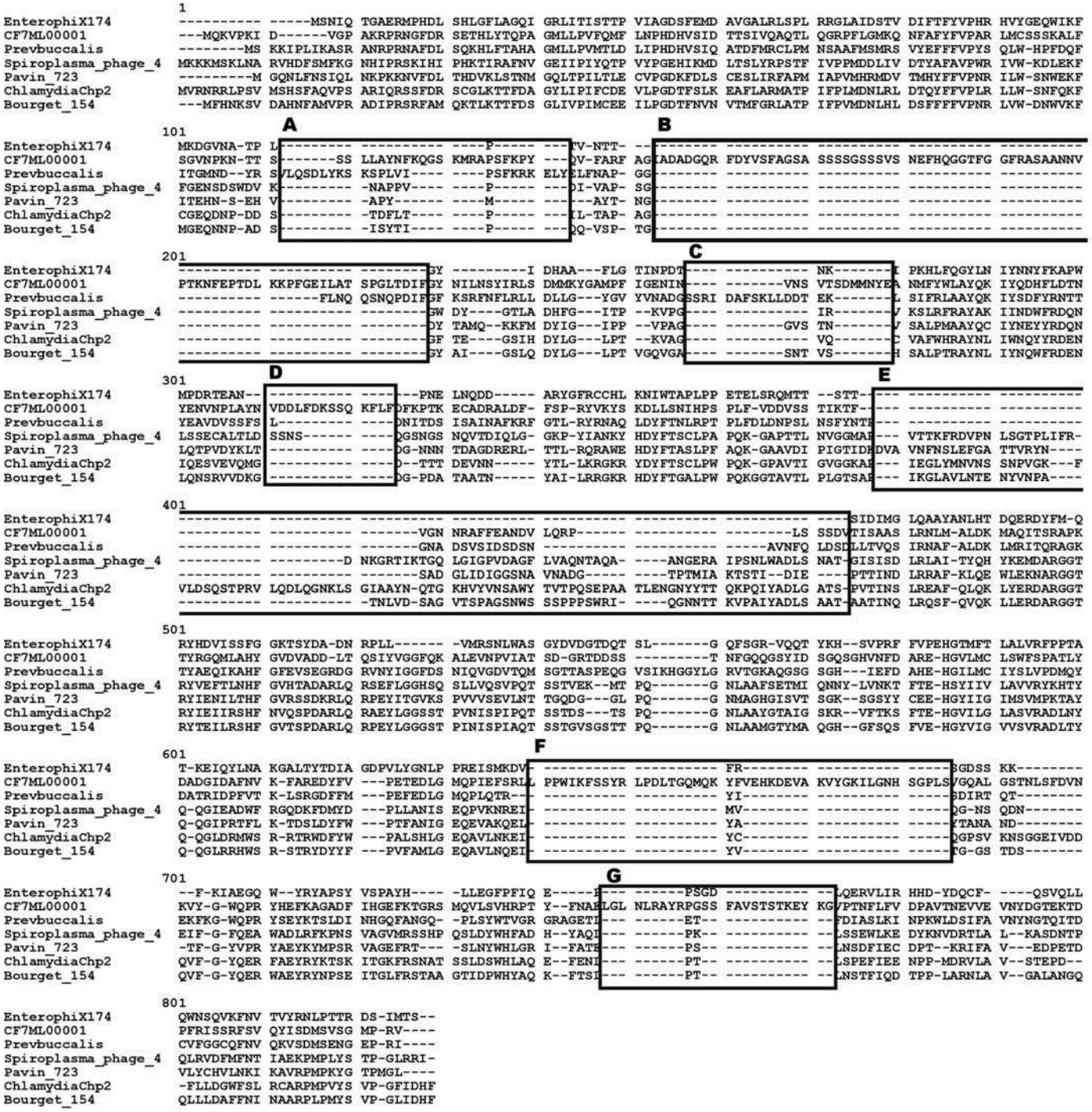


Figure S3 : Multiple amino acid alignment of the major capsid protein. Large insertions (more than 10 aa) are framed and identified from A to G. The insertion retrieved in all *Microviridae* but *Enterobacteria* phages known to induce mushroom-like structure is identified as the insertion E. One or several sequences were taken for each group, φX174 for *Enterobacteria* phages, CF7ML00001 and *Prevotella Buccalis* for *Alpavirinae,* Pavin_00723 for *Pichovirinae*, *Chlamydia* phage Chp2 and Bourget_00154 for *Gokushovirinae* and *Spiroplasma* phage 4.

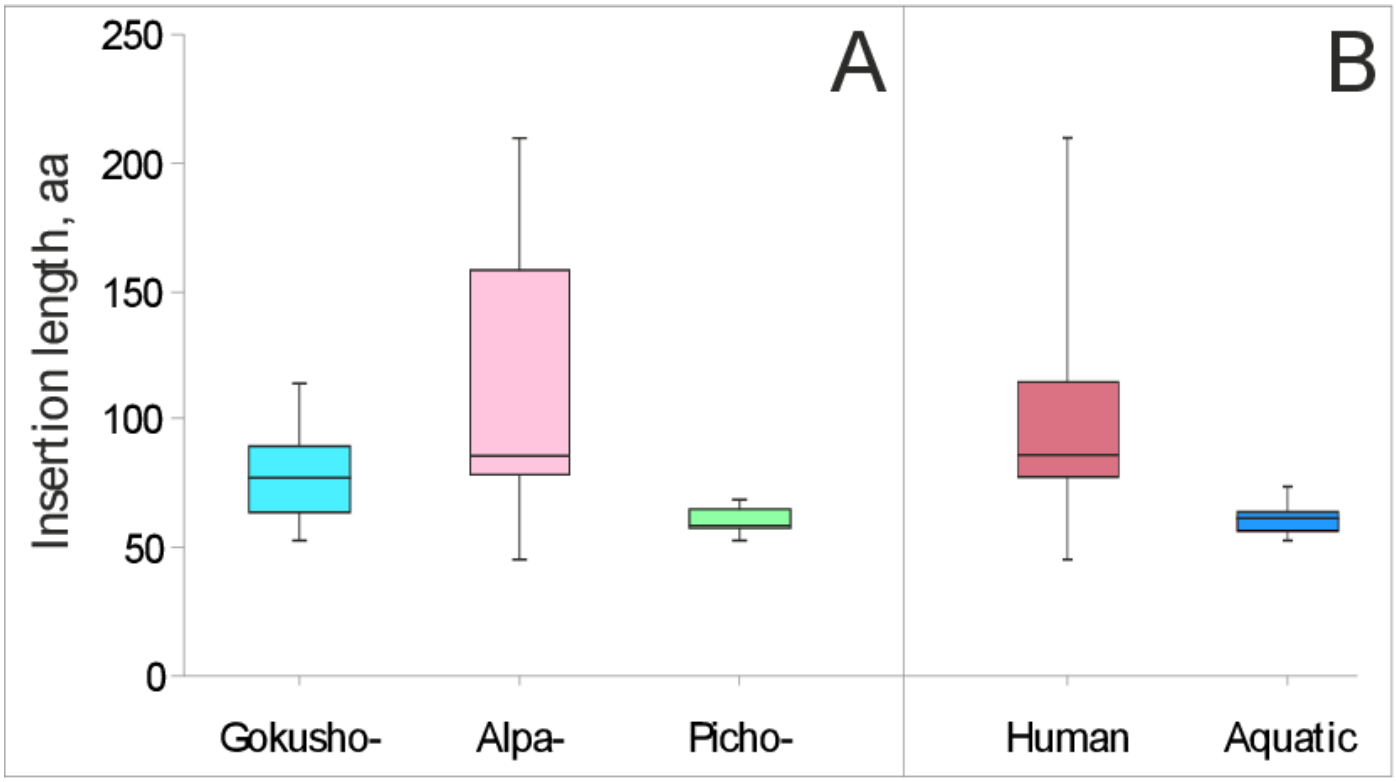


Figure S4 : A boxplot illustrating length variation of the 'mushroom-like' protrusion-forming insertions in the major capsid proteins of *Gokushovirinae*, *Alpavirinae*, and *Pichovirinae*. The insertion lengths are plotted as a function of the *Microviridae* subgroup (A) and ecosystem type (B).

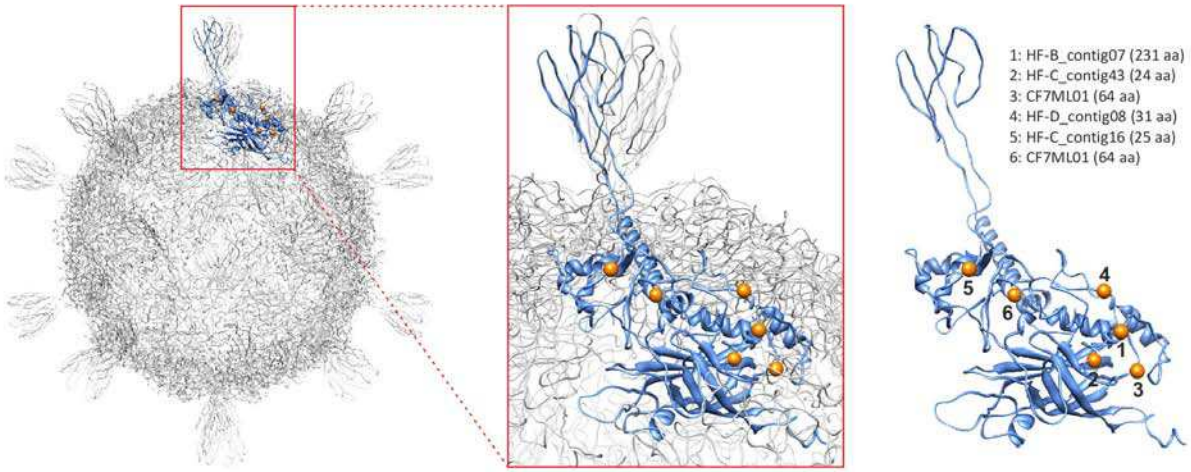


Figure S5 : Alpaviral VP1 in the context of the entire virion. Pseudoatomic model of the gokushovirus SpV4 virion (PDB ID:1KVP) with one of the capsomers substituted with the structural model of the alpaviral VP1 (*Prevotella bucalis* prophage BMV5). The hot-spots in the alpaviral VP1s where specific insertions (>15 aa) with respect to the BMV5 VP1 sequence were detected are indicated with orange spheres. The length of the largest insertion at each of the hot-spots is indicated along with the name of a corresponding viral genome. HF, human feces.

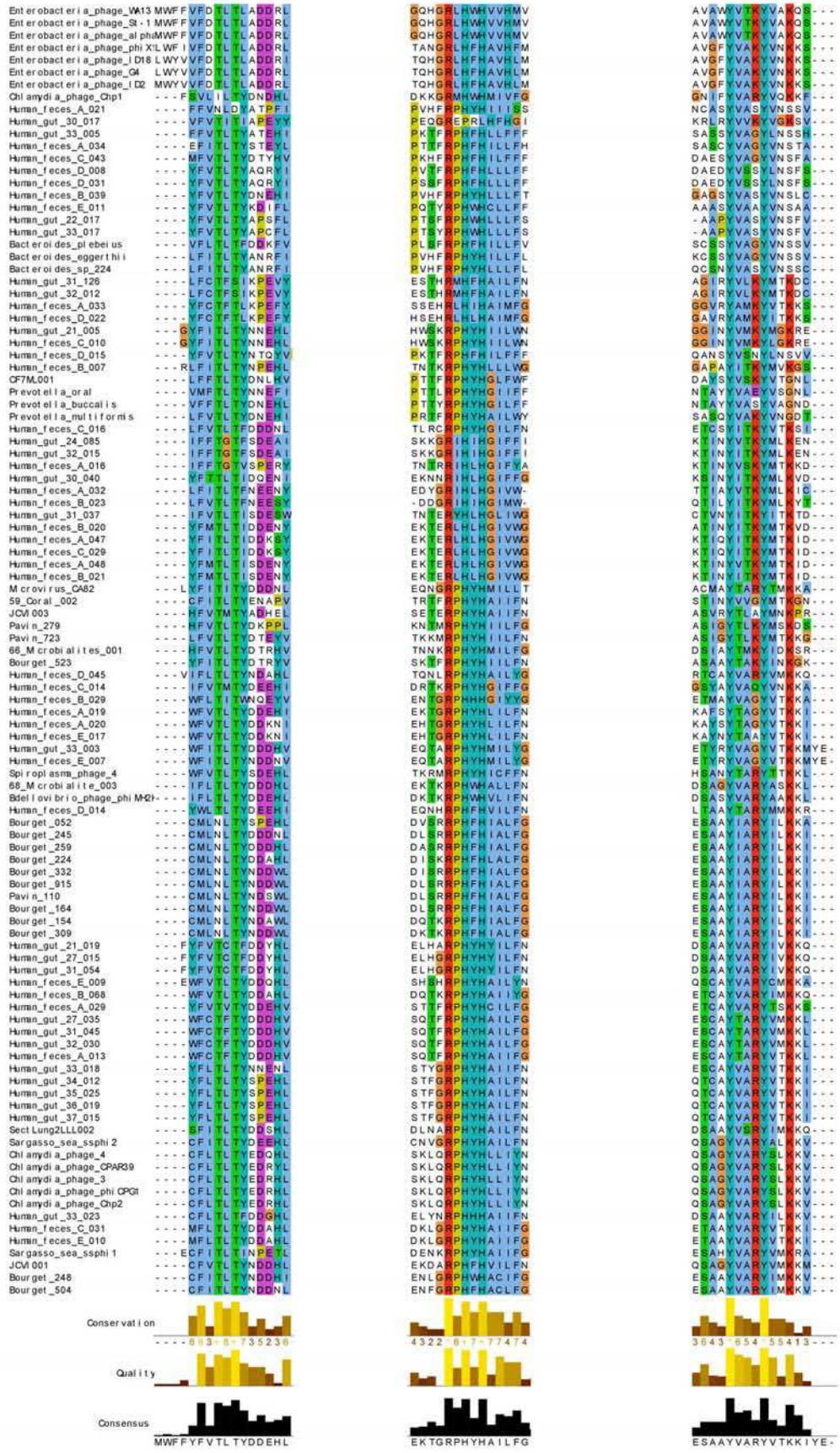


Figure S6 : Alignment of the conserved motifs of the superfamily I rolling-circle replication protein.

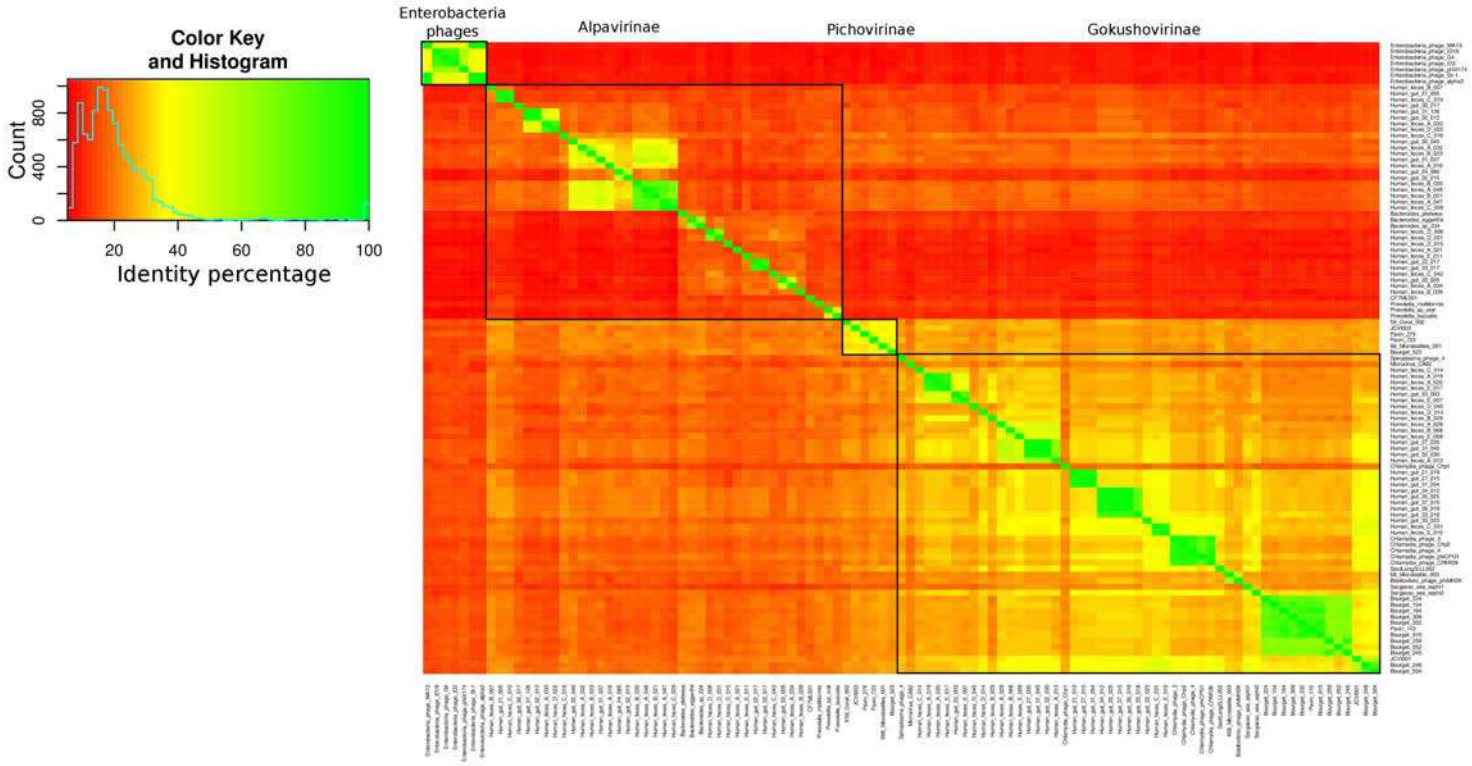


Figure S7 : Heatmap based on the percentage of identity from the replication protein multiple alignment. Scale is indicated on the top left, with the distribution of the percentages of identity. The genome affiliation is indicated above the heatmap, and groups are framed on the heatmap.

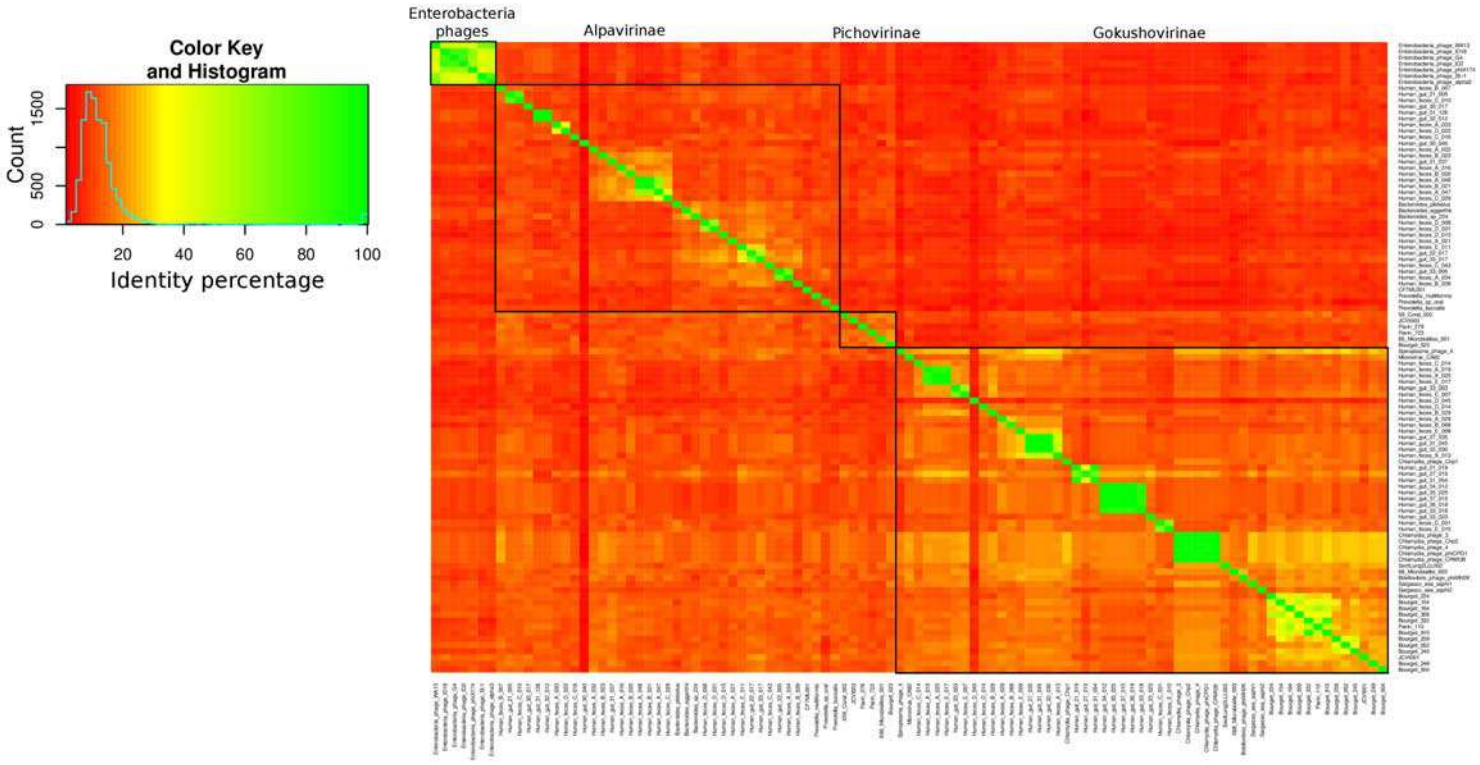


Figure S8 : Heatmap based on the percentage of identity detected on the capsid assembly protein multiple alignment. Scale is indicated on the top left, with the distribution of the percentages of identity. The genome affiliation is indicated above the heatmap, and groups are framed on the heatmap.

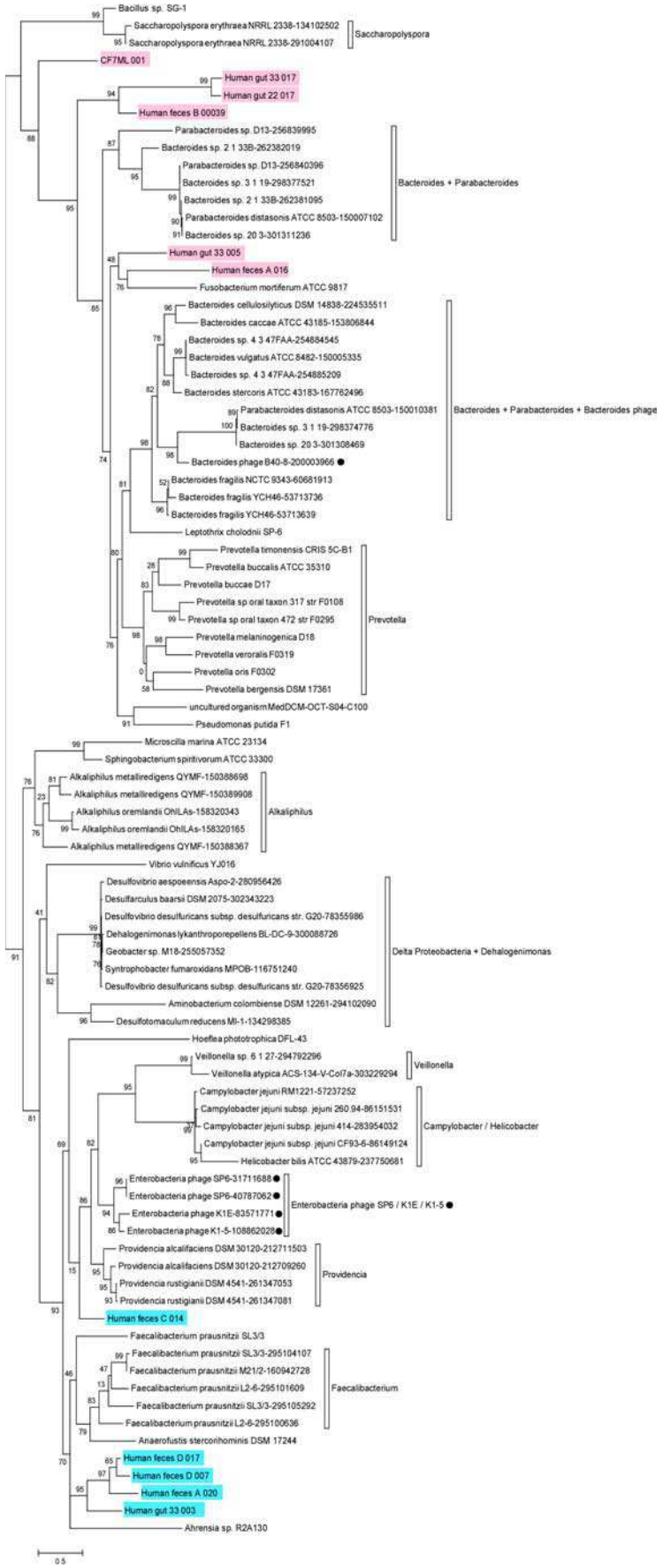


Figure S9 : Maximum-likelihood phylogenetic tree based on peptidase M15_3 protein sequences. Each reference sequences is identified by its name, followed by its gene id. *Alpavirinae* sequences are highlighted in pink, *Gokushovirinae* in blue, and viral reference sequences are marked with a black circle.

Annexe A.8 : Matériel supplémentaire

Supplementary material : Chimeric viruses blur the borders between the major groups of eukaryotic single-stranded DNA viruses (Article VII)

SI MATERIALS AND METHODS

Structural modeling and model quality assessment

The three dimensional model of the putative capsid protein (CP) of CHIV10 was constructed using the advanced multi-template approach using MODELLER v9.9 [1]. X-ray structures of tomato bushy stunt virus (TBSV; PDB ID: 2TBV) [2], melon necrotic spot virus (MNSV; PDB ID: 2ZAH) [3], carnation mottle virus (CMV; PDB ID: 1OPO) [4] and turnip crinkle virus (TCV; PDB ID: 3ZXA) [5] were used as templates. Sequence of CHIV10 CP was aligned with the corresponding sequences of TBSV, MNSV, CMV and TCV and the resultant alignment used to build a three-dimensional model of the putative CP of CHIV10. The initial model was optimized via multiple rounds of loop refinement with MODELLER. The stereochemical quality of the model was then assessed with ProSA-web [6]. ProSA-web quality (Z) score for the CHIV10 model was calculated to be -6.49, which is similar to the Z scores determined for the template structures (TBSV, -5.18; MNSV, -6.26; CMV, -6.06; TCV, -3.39) (Fig. SX). The MNSV virion map was downloaded from the VIPER database (viperdb.scripps.edu/) and rendered using UCSF Chimera [7].

Phylogenetic trees computation

Multiple alignments for RC-Rep and capsid genes were computed with MUSCLE [8], and manually curated. Maximum-likelihood phylogenetic trees were computed with FastTree [9] and annotated with Itol [10]. Additional chimeric-like capsid sequences were extracted from Lake Needwood RNA virome (44), based on BLASTp comparisons with capsid genes from the chimeric contigs assembled in this study and from BSL_RDHV genome [11].

Estimation of evolutionary distances between proteins

MEGA 5 [12] was used to assess evolutionary distances between protein sequences of capsid and RC-Rep genes (JTT model, gamma parameter set to the default value of 1.3). For ssDNA and ssRNA viruses, all available genomes were downloaded from NCBI, and clustered based on taxonomy (one genome for each species) and on global sequence similarity (threshold of 75% identity) with Uclust [13]. Comparisons were made within each taxonomic group *(Circoviridae, Geminiviridae, Nanoviridae* and *Tombusviridae*) and between chimeric viruses based on distinct multiple alignments computed with MUSCLE [8]. In order to keep the chart clear and viewable, only distances below 25 were taken, which removed 30 comparisons between *Geminiviridae* where RC-rep protein distances were below 10 but capsid genes distances were between 25 and 100. The same set of sequences was used in the genome size boxplot. Unclassified ssDNA and ssRNA viruses were not included in these analyses.

SUPPLEMENTARY REFERENCES

1 Marti-Renom, M. A. *et al.* Comparative protein structure modeling of genes and genomes. *Annu Rev Biophys Biomol Struct* **29**, 291-325 (2000).

2 Hopper, P., Harrison, S. C. & Sauer, R. T. Structure of tomato bushy stunt virus. V. Coat protein sequence determination and its structural implications. *J Mol Biol* **177**, 701-713 (1984).

3 Wada, Y. *et al.* The structure of melon necrotic spot virus determined at 2.8 A resolution. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **64**, 8-13 (2008).

4 Morgunova, E. *et al.* The atomic structure of Carnation Mottle Virus capsid protein. *FEBS Lett* **338**, 267-271 (1994).

5 Hogle, J. M., Maeda, A. & Harrison, S. C. Structure and assembly of turnip crinkle virus. I. X-ray crystallographic structure analysis at 3.2 A resolution. *J Mol Biol* **191**, 625-638 (1986).

6 Wiederstein, M. & Sippl, M. J. ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Res* **35**, W407-410 (2007).

7 Pettersen, E. F. *et al.* UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem* **25**, 1605-1612 (2004).

8 Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**, 1792-1797 (2004).

9 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490 (2010).

10 Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-478 (2011).

11 Diemer, G. S. & Stedman, K. M. A novel virus genome discovered in an extreme environment suggests recombination between unrelated groups of RNA and DNA viruses. *Biol Direct* **7**, 13 (2012).

12 Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**, 2731-2739 (2011).

13 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).

SUPPLEMENTARY FIGURES

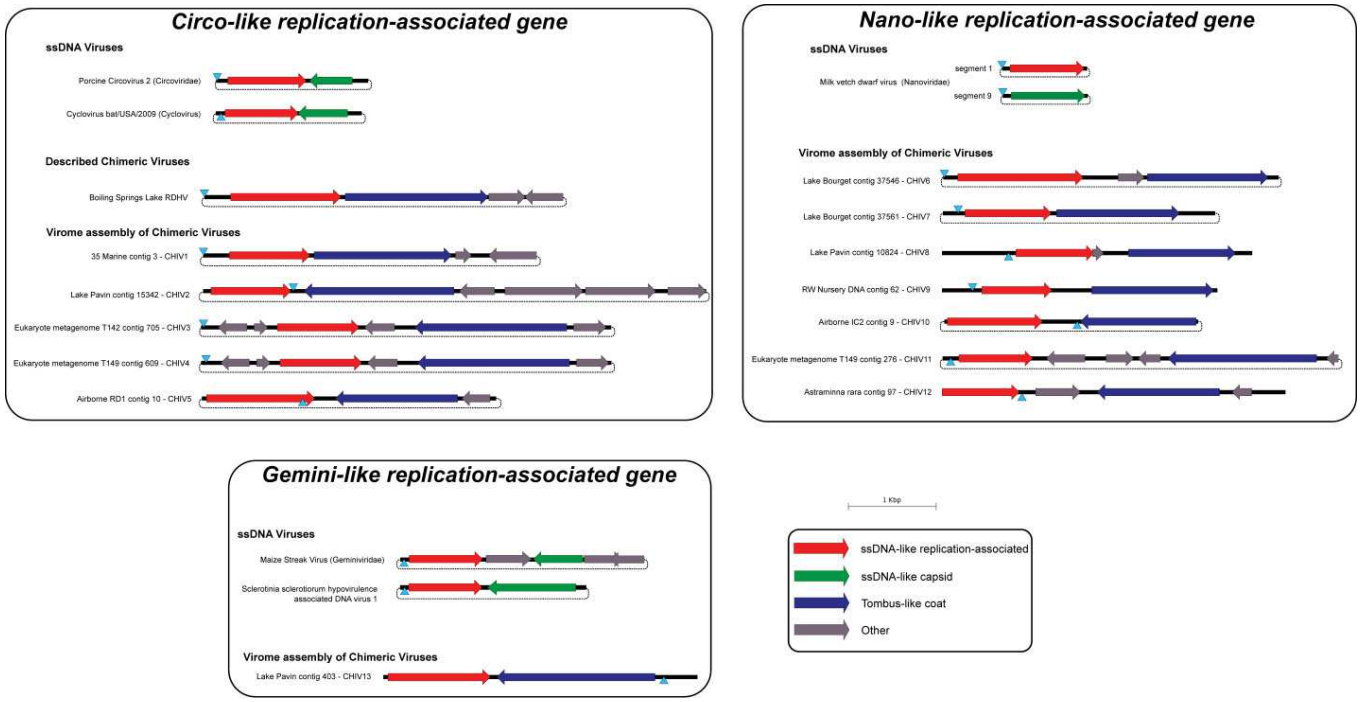


Figure S1. Genomic maps of Chimeric viruses and representative reference genomes. CHIV13 genome is reverse-complemented in order to start with the RC-Rep gene.

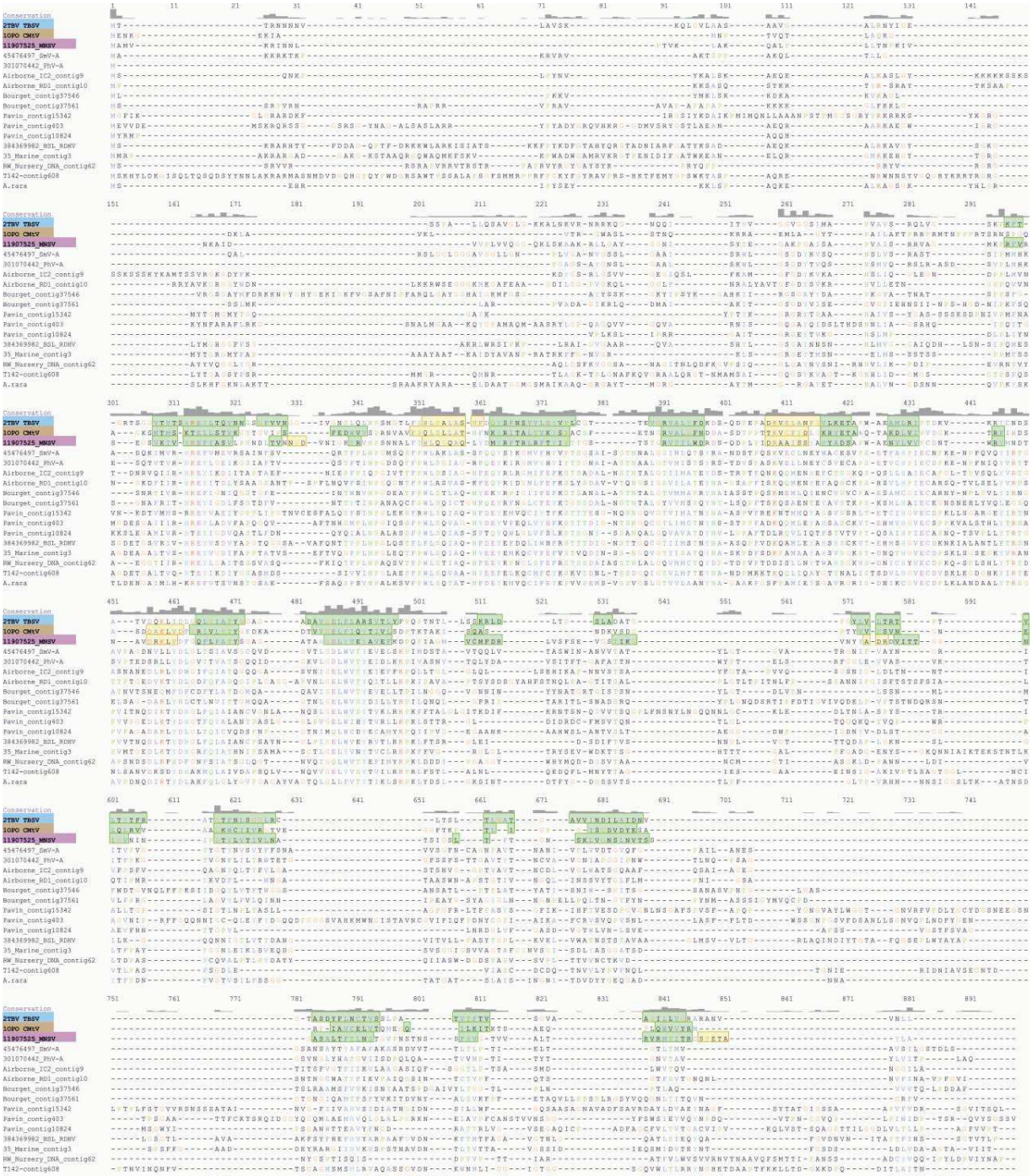


Figure S2. Alignment of capsid protein of CHIVs and representative tombusvirus references.

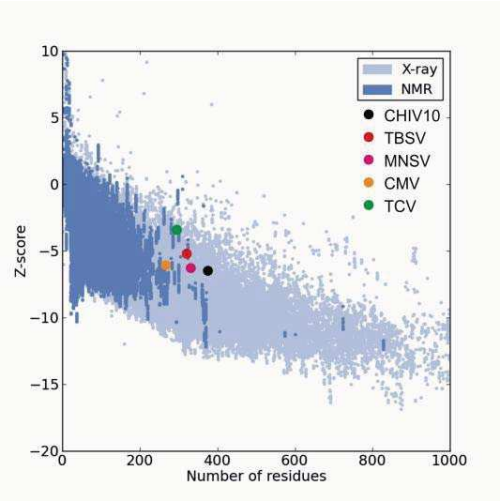


Figure S3. Quality assessment of the three-dimensional model of the CHIV10 CP. Quality of the generated model along with that of structural homologues used for modeling was evaluated using PsoSA-web at https://prosa.services.came.sbg.ac.at/prosa.php. The calculated quality (Z) scores (closed circles) are displayed in the context of the Z-scores of all experimentally determined protein structures available in the Protein Data Bank. Every dot represents a distinct structure solved by X-ray crystallography (light blue) or NMR (dark blue).

SUPPLEMENTARY TABLES

Table S1. List of DNA viromes screened for the presence of chimeric viruses, and RNA viromes searched for homologs of Tombus-like capsid genes.

Virome name	Available on	Methodology used in sample preparation	Sample origin	Sample type	Number of sequences	Average size of sequences	Number of contigs displaying similarities to a Tombus-like capsid gene	Number of contigs representing putative complete CHIV genomes
12_Saline_MedSalternSDbayVir112805	MG-Rast – 4440427.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	39,439	100.43		
13_Saline_MedSalternSDbayVir111605	MG-Rast – 4440428.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	58,319	98.11		
14_Saline_HighSalternSDbayVir111605	MG-Rast – 4440421.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	151,180	99.76		
15_Saline_LowSalternSDbayVir0704	MG-Rast – 4440436.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	268,049	104.47		
16_Saline_LowSalternSDbayVir111005	MG-Rast – 4440432.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	109,836	104.35		
17_Saline_MedSalternSDbayVir111005	MG-Rast – 4440431.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	39,348	101.40		
18_Saline_MedSalternSDbayVir112205	MG-Rast – 4440417.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	55,142	100.79		
19_Saline_HighSalternSDbayVir120705	MG-Rast – 4440145.4	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	46,628	102.15		
20_Saline_HighSalternSDbayVir112805	MG-Rast – 4440144.4	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	4,536	100.15		
21_Saline_LowSalternSDbayVir112805	MG-Rast – 4440420.3	PEG – CsCL- MDA	Solar Salterns – California	Hypersaline	62,363	103.87		
22_Saline_SaltonSeaVirOne082308	MG-Rast – 4440327.3		Salton Sea – California	Hypersaline	55,467	103.60		
23_Saline_SaltonSeaVirTwo082308	MG-Rast – 4440328.3		Salton Sea – California	Hypersaline	29,814	99.20		
32_Marine_GOMVir94to01	MG-Rast – 4440304.3	CsCL – MDA	Gulf of Mexico	Seawater	262,501	101.59		
33_Marine_BBCVir96to04	MG-Rast – 4440305.3	CsCL – MDA	British Columbia	Seawater	414,964	102.14		
34_Marine_ArcticVir2002	MG-Rast – 4440306.3	CsCL – MDA	Arctic sea	Seawater	686,209	99.15		
35_Marine_SARVir063005	MG-Rast – 4440322.3	CsCL – MDA	Sargasso sea	Seawater	397,939	104.31	1	1
36_CoralAtol_KingLIVir082105	MG-Rast – 4440036.3	CsCL – MDA	Kingmann - Line Islands	Seawater	93,744	108.34		
37_CoralAtol_XmasLIVir080505	MG-Rast – 4440038.3	CsCL – MDA	Christmas - Line Islands	Seawater	279,882	110.56		
38_CoralAtol_PalmLIVir081805	MG-Rast – 4440040.3	CsCL – MDA	Palmyra - Line Islands	Seawater	318,178	104.78		
39_CoralAtol_FannLIVir081105	MG-Rast – 4440280.3	CsCL – MDA	Tabuaeran - Line Islands	Seawater	378,475	104.13		
40_MarineBay_4440102.3	MG-Rast – 4440102.3	PEG – CsCL- MDA	Tampa Bay – Florida	Seawater	279,129	103.95		
41_MarineBay_SkanBayAKVir092706	MG-Rast – 4440330.3		Skan Bay – Alaska	Seawater	30,831	104.56		
46_Fresh_TilPondKentSTVir0806	MG-Rast – 4440424.3	PEG – CsCL- MDA	Tilapia Pond 3 – California	Freshwater	56,549	101.06		
47_Fresh_TilPondKentSTVir050406	MG-Rast – 4440412.3	PEG – CsCL- MDA	Healthy fish Pond – California	Freshwater	60,135	101.07		
48_Fresh_PrePondKentSTVir050406	MG-Rast – 4440414.3	PEG – CsCL- MDA	Prebead Pond – California	Freshwater	67,785	103.21		
49_Fresh_TpondKentSTVir1105	MG-Rast – 4440439.3	PEG – CsCL- MDA	Tilapia Pond – California	Freshwater	264,844	102.25		
57_Coral_T0PortComHawVir022306	MG-Rast – 4440376.3	CsCL – MDA	Porites Compressa	Eukaryote	39,113	101.32		
58_Coral_ConPorCompHawVir0206	MG-Rast – 4440374.3	CsCL – MDA	Porites Compressa	Eukaryote	39,191	103.70		
59_Coral_DOCPorCompVirHaw0206	MG-Rast – 4440370.3	CsCL – MDA	Porites Compressa	Eukaryote	35,409	102.18		
60_Coral_pHPorCompHawVir0206	MG-Rast – 4440371.3	CsCL – MDA	Porites Compressa	Eukaryote	49,949	104.73		
61_Coral_NutPorCompHawVir0206	MG-Rast – 4440377.3	CsCL – MDA	Porites Compressa	Eukaryote	34,139	107.18		
62_Coral_TempPorCompHawVir0206	MG-Rast – 4440375.3	CsCL – MDA	Porites Compressa	Eukaryote	38,482	113.38		
66_Microbialites_PAStromCCMexVir072205	MG-Rast – 4440320.3	CsCL – MDA	Paztac Azules	Microbialites	301,264	104.64		
67_Microbialites_RMStromCCMexVir072205	MG-Rast – 4440321.3	CsCL – MDA	Rio Mesquites	Microbialites	324,500	104.22		

68_Microbialites_HBCStromBahamasVir011105	MG-Rast – 4440323.3	CsCL – MDA	Bahamas	Microbialites	148,334	100.52		
73_Fish_FishHealSlimKentSTVir050406	MG-Rast – 4440065.3		Healthy fish gut	Eukaryote	61,022	98.45		
74_Fish_FishMorSlimKentSTVir050406	MG-Rast – 4440064.3		Morbid fish gut	Eukaryote	59,599	98.32		
84_Animal_HealSputSDRep3Vir070706	MG-Rast – 4440442.4	CsCL – MDA	Human Lung (USA)	Eukaryote	39,489	84.80		
85_Mosquito_MosqISDVir01252006	MG-Rast – 4440052.3	CsCL – MDA	Mosquito (USA)	Eukaryote	336,760	102.61		
86_Mosquito_MosqDigSDVir060606	MG-Rast – 4440053.3	CsCL – MDA	Mosquito (USA)	Eukaryote	638,689	100.32		
87_Mosquito_MosqIISDVir060606	MG-Rast – 4440054.3	CsCL – MDA	Mosquito (USA)	Eukaryote	601,040	104.16		
Antarctic_Lake_LopezBueno_Spring	Metavir – Project Lake Limnopolar	Sucrose cushion – MDA	Lake Limnopolar	Freshwater	41,322	239.65		
Antarctic_Lake_LopezBueno_Summer	Metavir – Project Lake Limnopolar	Sucrose cushion – MDA	Lake Limnopolar	Freshwater	38,475	221.27		
CF10LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	188,287	217.42		
CF6LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	156,809	217.66		
CF7LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	184,666	228.15		
CF8LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	158,912	246.84		
CF9LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	233,854	205.68		
Human_Salivary_Sub1_Day1	Metavir – Project Human_Salivary	CsCL – MDA	Human Salivary (USA)	Eukaryote	63,476	0.00		
Human_Salivary_Sub1_Day30	Metavir – Project Human_Salivary	CsCL – MDA	Human Salivary (USA)	Eukaryote	86,362	0.00		
Human_Salivary_Sub2_Day30	Metavir – Project Human_Salivary	CsCL – MDA	Human Salivary (USA)	Eukaryote	119,621	0.00		
Human_Salivary_Sub3_Day30	Metavir – Project Human_Salivary	CsCL – MDA	Human Salivary (USA)	Eukaryote	103,744	0.00		
Human_Salivary_Sub5_Day1	Metavir – Project Human_Salivary	CsCL – MDA	Human Salivary (USA)	Eukaryote	604,957	0.00		
Lake_Bourget	Metavir – Project French Lakes	PEG – MDA	Lake Pavin – France	Freshwater	649,290	412.31	4	2
Lake_Pavin	Metavir – Project French Lakes	PEG – MDA	Lake Bourget – France	Freshwater	593,084	433.41	6	3
GOS_Mv_858	Metavir – Virome GOS Move858	.22 µm filtration	Chesapeake Bay – Maryland	Seawater	11,496	1014.07		
Norm3LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	245,109	220.94		
Norm4LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	235,755	212.00		
Norm5LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	211,401	226.10		
Norm6LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	293,359	219.66		
Norm7LungVir20080407_nett	Ncbi – BioProject 39545	CsCL – MDA	Human Lung (USA)	Eukaryote	284,384	210.21		
Potable_DNA	Metavir – Project Reclaimed Water	PEG-CsCL-MDA	Reclaimed Water (Florida, USA)	Freshwater	240,259	209.21		
RW_Effluent_DNA	Metavir – Project Reclaimed Water	PEG-CsCL-MDA	Reclaimed Water (Florida, USA)	Freshwater	262,097	227.87		
RW_Nursery_DNA	Metavir – Project Reclaimed Water	PEG-CsCL-MDA	Reclaimed Water (Florida, USA)	Freshwater	283,753	225.69	1	1
RW_Park_DNA	Metavir – Project Reclaimed Water	PEG-CsCL-MDA	Reclaimed Water (Florida, USA)	Freshwater	202,436	97.55		
SectLung2LLL-PVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	8,981	391.39		
SectLung2LMLVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	14,037	351.82		
SectLung2LULVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	14,559	398.31		
SectLung2RLLVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	9,232	387.51		
SectLung2RMLVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	5,882	304.36		
SectLung2RULVir20090504	Ncbi – BioProject 66313	CsCL – MDA	Human Lung (USA)	Eukaryote	9,026	368.91		
Gut_X1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – X1	Eukaryote	30,873	372.90		
Gut_L1_8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L1	Eukaryote	132,569	379.12		

Gut_L2_1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	61,104	370.82		
Gut_L2_7	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	148,781	370.31		
Gut_L2_8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L2	Eukaryote	16,955	375.02		
Gut_H1_7	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	13,048	378.90		
Gut_H1_8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	16,747	365.41		
Gut_H2_8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H2	Eukaryote	16,137	366.71		
Gut_L1_1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L1	Eukaryote	107,259	405.06		
Gut_H1_2	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	107,993	409.48		
Gut_H1_1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H1	Eukaryote	25,648	377.67		
Gut_H2_1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – H2	Eukaryote	23,614	372.91		
Gut_L3_1	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	33,489	369.66		
Gut_L3_2	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	76,090	384.09		
Gut_L3_7	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	15,166	382.16		
Gut_L3_8	Metavir – Project Human Gut Diet	CsCL – MDA	Human Gut (USA) – L3	Eukaryote	59,155	382.72		
Gut_F-A	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	113,054	431.05		
Gut_F-B	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	109,569	435.21		
Gut_F-C	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	68,391	437.07		
Gut_F-D	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	115,121	433.08		
Gut_F-E	Metavir – Project Human Feces	CsCL – MDA	Human Gut (South Korea)	Eukaryote	98,511	417.05		
Air_RD-1	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Residential district (South Korea)	Atmosphere	8,083	548.76	2	1
Air_RD-2	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Residential district (South Korea)	Atmosphere	16,638	839.41	1	
Air_FR-1	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Forest (South Korea)	Atmosphere	10,291	540.62		
Air_FR-2	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Forest (South Korea)	Atmosphere	13,025	546.57		
Air_IC-1	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Industrial Complex (South Korea)	Atmosphere	8,873	535.47		
Air_IC-2	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Industrial Complex (South Korea)	Atmosphere	11,705	541.20	1	1
Air_RD-Rain-1	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Residential district (South Korea)	Atmosphere	8,744	538.35		
Air_RD-Rain-2	Metavir – Project Airborne viruses	TFF-Ultrafiltration	Residential district (South Korea)	Atmosphere	10,085	545.12		
Sediment_Deep-sea_IO	Ncbi – BioProject 189588	PEG – CsCl- MDA	Ogosawara (Pacific Ocean)	Seawater sediment	58,952	308.15		
Sediment_Deep-sea_M	Ncbi – BioProject 189588	PEG – CsCl- MDA	Mariana (Pacific Ocean)	Seawater sediment	55,471	275.12		
Sediment_Deep-sea_S	Ncbi – BioProject 189588	PEG – CsCl- MDA	Shimokita (Pacific Ocean)	Seawater sediment	99,457	306.17		
						Total	16	9

RNA viromes							**Number of contigs containing a Chimeric-like capsid gene**	**Number of contigs representing putative complete CHIV genomes**
Lake_Needwood_Jun	Metavir – Project RNA_Lake_Needwood	TFF	Lake Djikeng (USA)	Freshwater	10,110	271.47	1	0
Lake_Needwood_Nov	Metavir – Project RNA_Lake_Needwood	TFF	Lake Djikeng (USA)	Freshwater	15,329	257.89	29	0
RW_Effluent_RNA	Metavir – Project Reclaimed Water	PEG-CsCL-MDA	Reclaimed Water (Florida, USA)	Freshwater	247,586	209.83	0	0
RW_Nursery_RNA	Metavir – Project Reclaimed Water	PEG-CsCL-MDA	Reclaimed Water (Florida, USA)	Freshwater	295,823	218.88	0	0

Table S2. Characteristics of chimeric virus genomes.

	Structure	Length	RC-Rep type	RC-Rep vs Capsid orientation	Number of additional genes	Nonanucleotide + ITR coordinates	Nonanucleotide sequence	Stem loop location	Stem loop strand
Reference ssDNA viruses									
Porcine circovirus 2 / *Circoviridae*	circular	1767	Circo	Reverse	0	10–20 | 31–40	TAGTATTAC	Before Rep	Same
Cyclovirus bat/USA/2009 / Cyclovirus	circular	1703	Circo	Reverse	0	59–73 | 85–99	TAGTATTAC	Before Rep	Reverse
Milk vetch dwarf virus (segment 1) / *Nanoviridae*	circular	1007	Nano	-	0	1–11 | 23–33	TAGTATTAC	Before Rep	Same
Maize streak virus / *Geminiviridae*	circular	2690	Gemini	Reverse	3	2512–2529 | 2542–2559	TAATATTAC	Before Rep	Reverse
Sclerotinia sclerotiorum hypovirulence associated DNA virus 1	circular	2166	Gemini	Reverse	0	2156–2163 | 9–16	TAATATTAT	Before Rep	Reverse
Described chimeric viruses									
BSL RDHV	circular	4100	Circo	Forward	2	1–17 | 29–45	AAGTATTAC	Before Rep	Same
Assembled chimeric viruses									
Seawater virome - 35 Marine contig3 - CHIV1	circular	3802	Circo	Forward	2	5–17 | 29–41	TAGTATTAC	Before Rep	Same
Freshwater virome - Lake Pavin contig15342 - CHIV2	circular	5733	Circo	Reverse	4	1028–1037 | 1066–1075	CTGTATTAC	After Rep	Same
Seawater eukaryote metagenome - Euk T142 contig705 – CHIV3	circular	4675	Circo	Reverse	4	27–34 | 68–75	TATTATTAC	Before Rep	Same
Seawater eukaryote metagenome – Euk T149 contig609 – CHIV4	circular	4677	Circo	Reverse	4	57–64 | 98–105	TATTATTAC	Before Rep	Same
Atmosphere virome - Airborne RD1 contig10 – CHIV5	circular	3354	Circo	Reverse	1	1149–1158 | 1170–1179	TAGTATTAC	3' End of Rep gene	Reverse
Freshwater virome – Lake Bourget contig37546 – CHIV6	circular	3824	Nano	Forward	1	13–20 | 32–39	TAGTATTAC	Before Rep	Same
Freshwater virome – Lake Bourget contig37561 – CHIV7	circular	3106	Nano	Forward	0	122–132 | 160–170	TGTTATTCC	Before Rep	Same
Freshwater virome – Lake Pavin contig10824 – CHIV8	linear	3536	Nano	Forward	1	757–768 | 799–810	AACTATTAT	Before Rep	Reverse
Freshwater virome – RW Nursery DNA contig62 – CHIV9	linear	3139	Nano	Forward	0	1575–1586 | 1609–1620	TAATGTTAC	After Rep	Same
Atmosphere virome – Airborne IC2 contig9 – CHIV10	circular	2892	Nano	Reverse	0	1514–1526 | 1541–1553	GTTTATTAC	After Rep	Reverse
Seawater eukaryote metagenome – Euk T149 contig276 – CHIV11	circular	4511	Nano	Reverse	3	94–102 | 123–131	TATTATTAC	Before Rep	Reverse
Foraminifera Whole Genome Sequencing – Astraminna rara contig 97 – CHIV12	linear	3915	Nano	Reverse	2	1616–1624 | 1637–1645	TAACATTAT	After Rep	Reverse
Freshwater virome – Lake Pavin contig403 – CHIV13	linear	3581	Gemini	Reverse	0	268–280 | 294–306	TAATGTTAT	Before Rep	Reverse

Table S3. Detection of RCR and SH3 motifs in CHIVs RC-Rep gene sequences.

	RCR Motifs			SF3 Helicase motifs		
	I	*II*	*III*	*Walker-A*	*Walker-B*	*Walker-C*
References						
Circoviridae & cycloviruses	[VC]FT[LIW]NN	[Px]HLQG	YC[Sx]K	G[Px][Pst][Gc]xGKS	[VI][IUML]DDF	UTS[Ne]
Nanovirus	[VCx]FT[Li]N[FYN]	xHUQG	Y[CAs]xK	G[Ps]xG[Gn]EGK[TS]	[VIW][UAC][Fim]D[IVF]	V[FMI][Ac]N
Geminiviridae	FLTY[Ps]x	[Px]H[Lx]H[VAC]	Y[Uac]xK	Gx[ST]R[Ti]GK[Ts]	[VI][IV]DD[VI]	UL[Cx]N
Chimeric viruses						
BSL RDHV	CITVNN	KHLQV	YCTK	GATGTGKS	IIDDY	ITCP
35 Marine contig3 – CHIV1	CFTLNN	RHLQG	YCSK	GPTGTGKT	IIDDY	ITAP
Lake Pavin contig15342 – CHIV2	VLVLNN	IHLQG	YITK	GETGQGKS	IIDDY	ITTP
Euk T142 contig705 – CHIV3	CITHNN	KHIQA	YCIK	GETGTGKS	IIDDV	FTAP
Euk T149 contig609 – CHIV4	CITHNN	KHIQA	YCIK	GETGTGKS	IIDDV	FTAP
Airborne RD1 contig10 – CHIV5	VFTLNN	PHIQG	YCSK	GDTGTGKS	VINDF	ITSS
Lake Bourget contig37546 – CHIV6	CFTYYY	PHLQS	YCKK	DPKGGNGKT	-	VFSN
Lake Bourget contig37561 – CHIV7	DFRLNQ	IHFQG	YASK	DVKGCQGKQ	IFLDL	VFTN
Lake Pavin contig10824 – CHIV8	MFDWRH	DHYQG	YASK	DPVGNNGKT	TLIDL	VFTN
RW Nursery DNA contig62 – CHIV9	LLTYKS	PHTHA	YLSK	DSKGNAGKS	IIFDI	VMSN
Airborne IC2 contig9 – CHIV10	CCVYDF	KHFQG	YTMK	DEKGNIGKT	YLIDM	IFTN
Euk T149 contig276 – CHIV11	SWTWNK	LHYQG	YCMK	EDTGCTGKS	YIFDI	VFSN
Astraminna rara contig 97 – CHIV12	CFTFNN	PHLQG	YCSK	DEVGQLGKT	-	VLSN
Lake Pavin contig403 – CHIV13	HLTYKT	DHTHF	YHKK	GSTNTGKT	VFDDM	FTSN